7th International Scientific Conference
Technics and Informatics in Education
Faculty of Technical Sciences, Čačak, Serbia, 25-27th May 2018

Session 3: Engineering Education and Practice

UDC: 519.2 :613

Health Care Analysis Using Statistics

Andrijana Pešić [1*], Vera Lazarević [1], Marija Đukić [2]
[1] University of Kragujevac, Faculty of Technical Sciences Čačak, Srbija
[2] Technical College, Čačak, Srbija
[*] andrijana90pesic@gmail.com

Abstract: *In this paper, the data of a total of 215 patients with nine registered variables were processed. By applying appropriate statistical analysis, it has been found that variables within the groups do not have a normal distribution. Therefore, nonparametric analytical techniques were used. The results showed that the amount of consumed alcoholic beverages per day depends on the sex and that the mean value of the rank of this mark is higher in the male gender. Furthermore, tests have shown that marital status and the number of consumed caffeine drinks per day are dependent variable. The dependence of the marital status and the number of hours of sleep on weekends has been established. The paper also examines the correlation between some variables. The tests showed the existence of a negative correlation between the variables of the level of education and the number of cigarette smokers per day. The obtained results indicate that the level of education affects smoking behavior, which means that the prevention program should be the most intensive in the secondary school.*

Keywords: *statistics; health care; nonparametric techniques; hypothesis testing.*

1. INTRODUCTION

Over the past decades, a great increase in the use of statistical methods has been documented for a wide range of medical journals, [1]. Developing new practice and improving care of patients is primary goal of health care research, [2]. Since the appearance of individual cases may show smaller or larger deviations from the average or typical, it is necessary to observe in large numbers, in mass, to reveal what is in their general and legal, [3]. Hypothesis testing is prevalent in quantitative research in the field of health care, [4]. Results of testing depend on: how participants are selected and treated; process of data measurement; elimination or reduction of bias; planning a visit; expectations of patients and researchers; treating unwanted events; problem management, [5]. When assumptions of parametric data are violated nonparametrc test are used, and they can be used to analyse alcohol consumption directly using the categories, but results tend to be more conservative than parametric tests, [6, 7].

Since youth is increasingly prone to bad habits, at the early age, disadvantages of such habits should be pointed out through prevention programs.

The aim of the research is to determine the possible existence of relationships between individual random variables, as well as their intensity.

2. RESEARCH METHODOLOGY

In this paper, nonparametric techniques are represented as research methods, because the data don't have a normal distribution within the observed groups. The following techniques have been used: χ^2 independence test, Mann-Whitney U test, Kruskal-Wallis test, and Spirman's correlation.

The data on which the analysis was performed are used at the Medical Faculty in Kragujevac, within the course Medical Statistics and Informatics. The analysis used a sample of 215 patients, with the following variables being registered: sex, marital status, education level, smoker, number of cigarettes per day, number of alcoholic beverages per day, number of caffeinated beverages per day, number of hours of sleep in working days and number of hours of sleep on the weekends. Data was processed using statistical program SPSS.

3. χ^2 INDEPENDENCE TEST

χ^2 test of independence is used to test the independence of two categorical variables. We are testing the null hypothesis H_0: Sex and smoking behavior are independent variables, against the alternative hypothesis that it is not so. If the value of the test statistic or significance given in the Asymp.Sig. (2-sided) column is greater than the significance threshold (0.05), we have no reason to reject the null hypothesis and decide that these two features are independent, that is, the result is significant with the threshold significance of 0.05, or 95% confidence level. Otherwise, we reject the null hypothesis, our result is not significant and we decide that the observed features are dependent.

In Table 1, we can see that 12.7% of women are smokers and 87.3% are not. In men, 13.5% are smokers, while 86.5% are not smokers.

Table 1. *Cross-tabulation of variables sex and smoker/non-smoker*

			Smoker		Total
			Yes	No	
Sex	female	Count	16	110	126
		% within sex	12.7%	87.3%	100%
		% within Smoker	57.1%	58.8%	58.6%
		% of Total	7.4%	51.2%	58.6%
	male	Count	12	77	89
		% within sex	13.5%	86.5%	100%
		% within Smoker	42.9%	41.2%	41.4%
		% of Total	5.6%	35.8%	41.4%
Total		Count	28	187	215
		% within sex	13%	87%	100%
		% within Smoker	100%	100%	100%
		% of Total	13%	87%	100%

Since we have a 2x2 table, we read the value of Continuity Correction (Yates Correction). The corrected value is 0.000 with a significant of 1.00> 0.05, so we can conclude that our result is significant (Table 2). Therefore, we accept the null hypothesis that sex and smoking behavior are an independent variable.

Table 2. *The results of χ^2 independence test of variables sex and smoking behavior*

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (1-sided)	
Pearson Chi-Square	0.028[a]	1	.866		
Continuity Correction	.000	1	1.000		
Likelihood Ratio	.028	1	.866		
Fisher's Exact Test					
Linear-by-Linear Association	.028	1	.867	1.000	.511
N of Valid Cases	215				
a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 11.59.					
b. Computed only for a 2x2 table.					

4. MAAN-WHITNEY U TEST

Mann-Whitney U test is a nonparametric alternative to the t-test of independent samples. The Mann-Whitney test compares the median group that converts to ranks, so the distribution is not relevant. If the values differ from each other, then in one group there will be higher ranks, and in the other smaller.

We tested the hypothesis of the equality of medial values in the sexes for each of the following variables, in particular: the number of caffeinated drinks per day, the number of hours of sleep per working days, the number of hours of sleep on weekends. In all cases, we obtained Asymp.Sig. (2-sided)> 0.05, indicating that we should accept the null hypothesis, that is, we consider that the median of observed observations is the same (there is no statistically significant difference between the median of observed variables in women and in men, at a confidence level of 95%).

In Table 3, a nonparametric Mann-Whitney report is presented with a different conclusion. We test the hypothesis that the medial values of the variable - the number of alcoholic drinks per day in men and women, are different. For more than 30 elements in the sample, SPSS calculates the amount of z-approximation, which is now a continuous random variable, while the variable - the number of alcohol beverages per day was discrete random variable. This approximation also includes correction due to the interconnection between the data. The first table Ranks displays descriptive group information, the mean rank (mean rank gender has a higher rank that corresponds to a higher value on an uninterrupted scale of a new variable), and sum of ranks.

In the second table, the z-approximation test statistics is -2.482 with a significance level of p = 0.013 <0.05, which means that the result is not statistically significant and we reject the null hypothesis. Medial values of the variable - the number of alcohol beverages per day, in men and women are different, [8].

Table 3. *Report from SPSS for Mann-Whitney U test for the variable: number of alcoholic drinks per day*

Ranks	Sex	N	Mean Rank	Sum of Ranks
Number of alcoholic drinks per day	Female	126	99.67	12558.00
	Male	89	119.80	10662.00
	Total	215		

Test statistics[a]	Number of alcoholic drinks per day
Mann-Whitney U	4557.000
Wilcoxon W	12558.000
Z	-2.482
Asymp. Sig. (2-tailed)	.013
a. Grouping Variable: Sex	

Median both groups (female and male sex) can be seen in Table 4.

Table 4. *Median both groups (female and male sex)*

Sex	N	Median
female	126	.1250
Male	89	1.0000
Total	215	1.0000

The size of the impact can be calculated using the following formula:r=z/√N=(-2.482)/√215=-0.169.

In this example, z=-2482, and N=215, r is therefore 0.169. This would be considered as a very small influence on Koen's criterion (0.1 = small impact).

5. KRUSKAL-WALLIS TEST

The Kruskal-Wallis test is a nonparametric alternative to oneway anova, and is used to compare the results of three or more groups. Like all the nonparametric tests, the Kruskal-Wallis test is not as powerful as one-way anova.

We test the null hypothesis H_0: There is a difference in the number of drunk caffeinated drinks compared to marital status, against an alternative hypothesis that is not so (Table 5).

Table 5. *SPSS report for Kruskal-Wallis test for the variable: number of caffeinated drinks per day*

Ranks	Marital status	N	Mean Rank
Number of caffeinated drinks per day	Single	41	81.13
	Married	151	114.48
	Divorced	17	108.32
	Widowed	6	127.58
	Total	215	
Test statistics[a]		Number of caffeinated drinks per day	
Chi-Square		10.185	
Df		3	
Asymp. Si		.017	
Kruskal Wallis Test			
Grouping Variable: Marital Status			

Median rankings show that most caffeinated drinks are consumed by respondents with marital status, widow.

Based on Table 5, we can see that the level of significance is less than 0.05, so we can conclude that the difference in the obtained values of the continuous variable between these four groups is significant, which means that we accept the null hypothesis.

We still do not know which groups differ statistically, so several subsequent Man-Vitney tests need to be done. In addition to all the pairs, it is necessary to do the Bonfferoni correction of alpha value (alpha value is divided by the number of planned tests 0,05 / 6 = 0.008).

Based on the analysis we concluded that the group of free and married group differ in the number of drunk caffeinated drinks during the day. The size of the impact is r = -3.102/14.66 = 0.211, which is at the boundary between small and medium impacts.

In this part we will carry out some more research using Kruskal-Wallis test.

We will test the null hypothesis H_0: There is a difference in the number of hours of sleep on weekends in relation to marital status, against an alternative hypothesis that there is no difference.

Table 6. *SPSS report for Kruskal-Wallis test for the variable: number of hours of sleep/weekend*

Ranks	Marital status	N	Mean Rank
Number of hours sleep on the weekends	Single	41	127.78
	Married	151	106.25
	Divorced	17	74.15
	Widowed	6	112.67
	Total	215	
Test statistics[a]		Number of hours sleep on the weekends	
Chi-Square		10.008	
df		3	
Asymp. Sig.		.019	
Kruskal Wallis Test			
Grouping Variable: Marital Status			

Based on the results of the tests shown in Table 6, we conclude that there is a difference in the values of the tested markings, depending on the marital status, because the level of significance is less than 0.05. This means that we have no reason to reject the null hypothesis, with a significance threshold of 0.05.

We will additionally make a comparison of groups free and married, and a group married and divorced. For both comparisons, the level of significance is less than 0.05 (0.029; 0.036 <0.05), which indicates that there is a difference in the number of hours of sleep on weekends in relation to the indicated groups.

In the following part we carried out analyzes on the sub-sample of the observed sample, which included all cigarette smokers. This sub-sample is divided into groups depending on the level of education (Table 7).

Table 7. *SPSS Report for Kruskal-Wallis test for variable: number of cigarettes per day*

Ranks	Education level	N	Mean Rank
Number of hours sleep on the weekends	Primary school	3	94
	Secondary school	23	131.80
	Post secondary school	26	115.25
	Undergraduate degree	60	108.13
	Postgraduate degree	103	101.18
	Total	215	

Test statistics[a]	Number of cigarettes per day
Chi-Square	14.940
df	4
Asymp. Sig.	.005
a. Kruskal Wallis Test	
a. Grouping Variable: Education level	

The significance level is 0.005<0.05, so we can conclude that there is a statistically significant difference between the groups compared to the number of cigarettes smoked. We can determine which groups differ more or less differently from the Mann-Whitney U test (Figures 8, 9 and 10).

Table 8. *SPSS report for Mann-Whitney test on the difference in ranking of variable secondary schools and undergraduate student in relation to the number of cigarettes per day*

Ranks	Education level	N	Mean Rank	Sum of Ranks
Number of cigarettes per day	Secondary school	23	48.65	1119.0
	Undergraduate degree	60	39.45	2367.00
	Total	83		

Test statistics[a]	Number of cigarettes per day
Mann-Whitney U	537.000
Wilcoxon W	2367.000
Z	-2.261
Asymp. Sig. (2-tailed)	.024
a. Grouping Variable: Education level	

Based on the Table 8, by examining the mean values of the ranks, we can see that the variable-secondary school has a higher rank corresponding to the higher value of the continuous variables. Test statistics show a statistically significant difference between the number of cigarette smoking groups of high school and undergraduate student, and this number is significantly higher for the first group.

Table 9. *SPSS Report for the Mann-Whitney test of the rank of variable secondary schools and postgraduate studies in relation to the number of cigarettes per day*

Ranks	Education level	N	Mean Rank	Sum of Ranks
Number of cigarettes per day	Secondary school	23	78.26	1800.00
	Postgraduate degree	103	60.20	6201.00
	Total	126		

Test statistics[a]	Number of cigarettes per day
Mann-Whitney U	845.000
Wilcoxon W	6201.000
Z	-3.813
Asymp. Sig. (2-tailed)	.000
a. Grouping Variable: Education level	

In Table 9, we can see the results of the Man-Vitney test that showed that there is a difference between groups of secondary schools and postgraduate studies in the number of smoked cigars.

Table 10. *SPSS report for Mann-Whitney test rank difference variables undergraduate and postgraduate studies in relation to the number of cigarettes per day*

Ranks	Education level	N	Mean Rank	Sum of Ranks
Number of cigarettes per day	Undergraduate degree	26	71.65	1863.00
	Postgraduate degree	103	63.32	6522.00
	Total	129		

Test statistics[a]	Number of cigarettes per day
Mann-Whitney U	1166.000
Wilcoxon W	6522.000
Z	-2.016
Asymp. Sig. (2-tailed)	.044
a. Grouping Variable: Education level	

Also, the test statistic showed a statistically significant difference (0.044 <0.05) between undergraduate and postgraduate studies in relation to the number of cigarettes smoked, as shown in Table 10. In all of these above-mentioned results, the higher mean value of the rank means higher consumption of cigarettes.

6. SPEARMAN'S CORRELATION

"Correlation" as a popular term implies simply a relationship among events. It refers to a quantitative expression of the interrelationship, or association, namely a coefficient of correlation.

The level of association is measured by how tightly or loosely the (x,y) observations cluster about the line. Because this coefficient is standardized by dividing by the standard deviations, it lies in the range −1 to +1, with 0 representing no relationship at all and ±1 representing perfect predictability. A positive coefficient indicates that both variables tend to increase or decrease together, whereas with a negative coefficient, one tends to increase as the other decreases, [9].

Table 11. *Spirman's correlation between the level of education and the number of cigarettes per day*

Correlations			Education level	Number of cigarettes per day
Spearman's rho	Education level	Correlation Coefficient	1.000	-.220**
		Sig. 2-tailed	.	.001
		N	215	215
	Number of cigarettes per day	Correlation Coefficient	-.220**	1.000
		Sig. 2-tailed	.001	.
		N	215	215
**Correlation is significant at the 0.01 level (2-tailed)				

Based on the results shown in Table 11, we can conclude that there is a negative correlation (small) between the level of education and the number of smoked cigarettes. The negative correlation indicates that with increasing the level of education reduces the number of smoked cigarettes.

7. CONCLUSION

The need for extensive statistical research in the field of health care will be inevitable in the future to monitor the negative health habits that affect the quality of life of the population in order to present a program of prevention.

When creating a prevention program, statistical analyzes should be performed on large sample and it is necessary to collect data with as many characteristics as possible in order to get more accurate information.

The results obtained showed that there was a difference in the number of alcoholic drinks between men and women. Single and married differ in the number of caffeinated drinks per day. The number of cigarettes smoked differs depending on the level of education, and with the increase in the level of education, the number of cigarettes smoked is reduced. Group of secondary school is predominant in the smoking intensity of groups of undergraduate students and postgraduate studies.

It is necessary, from the beginning of the primary school, that children should indicate the harmfulness of smoking in order to develop awareness of what is good and what is bad. A more intensive prevention plan should be implemented in secondary school.

In order to achieve a better prevention plan it is necessary to take more characteristics some of the following during the examination, e.g. financial situation, number of children as well as their age, number of family members, place of residence... Continuous and planned population testing is the key to improving the habits of the population.

REFERENCES

[1] Strasak, A., Zaman, Q., Marinell, G., Pfeiffer, K. (2007). The Use of Statistics in Medical Research, *The American Statistician*, 47-55.

[2] Scott, I., Mazhindu, D. (2014). *Statistics for Healthcare Professionals: An Introduction,* United Kingdom: SAGE Publications Ltd.

[3] Vukadinović, S. (1973). *Elementi teorije verovatnoće i matematičke statistike*, Beograd: Privredni pregled.

[4] Barbara, H. M. (2005). *Statistical Methods for Health Care Research*, Philadelphia: Lippincott Williams & Wilkins.

[5] Harris, J., Boushey, C., Bruemmer, B., Archer, S. (2008). Publishing Nutrition Research: A Review of Nonparametric Methods, Part 3, *Journal of the American Dietetic Association*, 1488-1496.

[6] Elise, W., Ball, J. (2002). Statistics Review 6: Nonparametric Methods, *Critical Care*, 509–513.

[7] Peace, K., Parrillo, A., Hardy, C. (2008). Assessing the Validity of Statistical Inferences in Public Health Research: An Evidence-Based*, Journal of the Georgia Public Health Association*, 10-22.

[8] Janjić, M., Lazarević, V., Micić, Ž., Vujičić, M. (2013). Comparative analysis of students' high school results and mechanical engineering entrance exam*, 3rd International Conference on Information Society Technology and Management*, 3rd-6th March, Kopaonik, 221-226.

[9] Robert H. R. (2012). *Chapter 21 – Regression and Correlation in Statistics in Medicine (Third Edition),* Amsterdam:, Elsevier/Academic Press.