

Performing Chi-square Test to Identify Factors Related to Human Health in two rural zones of Albania

Erjola Cenaj^{*}, Raimonda Dervishi² Zhifka (Muka) Meço³

¹Department of Mathematical Engineering, Faculty of Mathematical and Physical Engineering, Polytechnic University of Tirana, Albania.

²Department of Mathematical Engineering, Faculty of Mathematical and Physical Engineering, Polytechnic University of Tirana, Albania.

³Department of Mathematical Engineering, Faculty of Mathematical and Physical Engineering, Polytechnic University of Tirana, Albania.

erjola_cenaj@yahoo.com

(Received: 23 August 2024, Accepted: 31 August 2024)

(5th International Conference on Engineering and Applied Natural Sciences ICEANS 2024, August 25-26, 2024)

ATIF/REFERENCE: Cenaj, E., Dervishi, R. & Meço, Z. M. (2024). Performing Chi-square Test to Identify Factors Related to Human Health in two rural zones of Albania. *International Journal of Advanced Natural Sciences and Engineering Researches*, 8(7), 414-417.

Abstract – In various research, the effects of air pollution on human health, causing various diseases in individuals, have been evidenced. This paper analyzes the data collected in two different rural zones in the state of Albania: one of which is polluted and the other is not. Moreover, different categorical variables are treated that are summarized as a series of numbers in a random table, called a contingency table. The category of individuals is divided into four age groups. The chi-square test of independence is used to determine if there is any difference between the study groups in the proportions of the existing risk factor. The result shows that at the significance level $\alpha=0.05$ of the test the relationship depends between the disease (cough) and air pollution. Considering the growing number of diseases in individuals as a result of air pollution, which is a challenge for the health system, we recommend that more emphasis be placed on the use of advanced statistical methods - in the search of human health by government researchers and health system experts.

Keywords – Air pollution; categorical variable, chi-square test, human health.

I. INTRODUCTION

Air pollution is an important risk factor for the health of individuals in Europe and worldwide. Exposure to high levels of air pollution causes various health consequences. The short-term and long-term public health consequences of exposure to air pollution must be carefully assessed, as they constitute a significant threat to public health [1]. The health effects of air pollution are well documented, although the mixtures of air pollution can be complex. Referring to recent studies of diseases, it is shown that it is one of the ten main risk factors for the health of individuals [2].

According to the World Health Organization report in 2019 about 6.7 million deaths worldwide, with almost 85% attributable to non-communicable diseases (NCDs), including ischemic heart disease, stroke, lung cancer, asthma, chronic obstructive pulmonary disease (COPD), and diabetes. This makes air pollution

the second leading cause of NCDs globally after tobacco. In the European Union, 400.000 people suffer from premature death [1].

Exposure to air pollution is closely related to acute and chronic health effects, ranging from irritant effects to death. Particularly affected are sensitive and vulnerable groups such as pregnant women, children, the elderly and those already suffering from respiratory and other serious illnesses or low-income groups [3].

This paper analyzes the data collected in two rural zones in the state of Albania: one of which is polluted and the other is not. Moreover, different categorical variables are treated that are summarized as a series of numbers in a random table, called a contingency table. The category of individuals is divided in four age groups. The chi-square test of independence is used to determine if there is any difference between the study groups in the proportions of the existing risk factor [4].

The Chi-square statistic is a non-parametric tool designed to analyze group differences when the dependent variable is measured at a nominal level. Like all nonparametric statistics, it is robust to the distribution of the data. It does not require equality of variances between the study groups or homoscedasticity in the data, furthermore allows the evaluation of both dichotomous independent variables and multiple studies in the group. Unlike many other non-parametric and some parametric statistics, the calculations required to calculate the Chi-square provide considerable information about how each of the groups performed in the study [5]. The result after performing Chi-test shows that at the significance level $\alpha=0.05$ of the test the relationship depends between the disease (cough) and air pollution. Furthermore, the disease (cough) depends on the AGE group for the selected two different zones.

II. DATA COLLECTED

Data collection was carried out from two different rural zones in Albania, one near Elbasan, a highly polluted air environment, and the other near Kukes, an environment with unpolluted air. The database is an Excel file contains age in years, presence or absence of the cough for 95 individuals: 55 from the polluted zone and 40 from the non-polluted one which were selected to participate in the study. Table 1 presents the number of individuals with the absence or presence of cough in the polluted and non-polluted zones is presented.

Table1. The cough data for the polluted and non-polluted zone.

Zone	Cough		
	Present	Absent	Total
Polluted Zone	39	16	55
Non-Polluted Zone	25	15	40
Total	64	31	95

It is of interest to explore the relationship between age and the presence or absence of cough in the polluted zone. For this, we group in intervals the age data in years. In Table 2 the AGE group variable and the frequency of occurrence of each cough value is presented.

Table 2. Frequency table of age group by disease(cough).

AGE Group	Cough		
	Present	Absent	Total
15-35	2	20	22
36-45	8	16	24
46-55	14	15	29
56-80	14	6	20
Total	38	57	95

III. METHODOLOGY

In this study we apply Chi-square test to identify factors related to human health in two rural zone of Albania: one of which is polluted and the other is not. It is an estimation method for categorical data which we will use to measure the distance between human health to the respective categories and the predicted expected number on diseases where the null hypothesis is true [6].

The null hypothesis assumes that the proportions of individuals with (or without) cough are the same for each zone, while the alternative hypothesis assumes that the proportions of individuals with (or without) cough are not the same for each zone. The Chi-Square test which is a nonparametric statistical approach can be used in experimental work for counting data and let us to investigate whether the proportions in different categories are the same for all populations.

We assumed also that the expected number of accidents is proportional to the ratio of corresponding categories in the population [7]. Then we calculate the value of the test statistic from the formula:

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{Observed} - \text{estimated Expeced})^2}{\text{estimated Expeced}}$$

Referring to the Chi-Square approach we find rejection region according to this rule: if the calculated value of χ^2 test statistic is greater than tabulated value (with 5% level of significant) than null hypothesis can be rejected, otherwise the null hypothesis is accepted.

IV. RESULTS

Initially in the study we are focused to test the dependence of the cough from the zone using the Chi-square statistic with 5% level of significance. The null hypothesis of homogeneity states that the proportions of individuals with (or without) cough are the same for each zone.

The data of Table 1 that presents the number of individuals with the absence or presence of the cough for the polluted and non-polluted zone, the calculated $\chi^2 = 5.201 \geq 3.843 = \chi^2_{0.05,1}$ show that H_0 can be rejected at 5% level of significance.

Table 2 shows the AGE group variable and the frequency of occurrence of each cough value. The hypothesis of homogeneity states there is no distinction between the AGE group and the presence or absence of cough.

From these data, the calculated $\chi^2 = 15.899 \geq 7.815 = \chi^2_{0.05,3}$ shows that H_0 can be rejected at 5% level of significance, showing that there is evidence that the cough depends on the AGE group.

V. DISCUSSION

Air pollution remains an important risk factor for the health of individuals in Europe and worldwide, accounting for about 6.7 million deaths worldwide in 2019 [8]. In the European Union 400,000 people suffer from premature death from air pollution [1].

This study belongs to two different rural zones in Albania, one near Elbasan, a highly polluted environment, and the other near Kukes, an environment without pollution. The database contains age in years, presence or absence of the cough for 95 individuals: 55 from the polluted zone and 40 from the non-polluted ones who were selected to participate in the study.

The scope of this study is to verify the relationship between the significant factors and human health. The study would be more effective if we had more categories of individuals available, however, we made a summary of cough disease in two different zones to better understand some of the important factors related to human health, concluding that the proportions of individuals with (or without) cough are not the same for each zone and the cough depends on the AGE group. Considering the growing number of diseases in individuals as a result of air pollution, which is a challenge for the health system, we recommend that more

emphasis be placed on the use of advanced statistical methods - in the search for human health by government researchers and health system experts.

VI. CONCLUSION

In this study we made a summary of cough disease on two different rural zones in Albania to better understand some of the important factors related to human health. The data are considered as categorical variables. We used Chi-square test which is widely used by medical professionals.

Based on statistical analysis it is found that the proportions of individuals with (or without) cough are not the same for each zone and at a significance level of $\alpha=0.05$.

REFERENCES

- [1] <https://www.who.int/news/item/25-06-2024-what-are-health-consequences-of-air-pollution-on-populations/>
- [2] European Environment Agency (2014), Air quality in Europe - 2014 report.
- [3] <http://env-health.org/>
- [4] <http://www.who.int/mediacentre/news/releases/2024/air-pollution/en/>
- [5] McHugh ML. The chi-square test of independence. *Biochem Med (Zagreb)*. 2013;23(2):143-9
- [6] Jean-Louis Auget N. Balakrishnan Mounir Mesbah Geert Molenberghs, (2007). *Advances in Statistical Methods for the Health Sciences*, ISBN-10: 0-8176-4368-0
- [7] A. Agresti, *Categorical Data Analysis*, 2nd ed., University of Florida, 2003, pp.36-78.
- [9] WHO, (2019). European Region, Fact sheets on sustainable development goals: health targets.