Statistical data analysis in the Statistics II lesson of the Department of Applied Informatics of the University of Macedonia in the years 1997-1999-Study and conclusions.

1 Abstract

The purpose of this project is to analyze the student's scores and to check their normality in the Statistics II lesson of June examinations in the periods from 1997 to 1999. The data will be analyzed before and after the removal of the writings, that were marked as zero. The results show that there is not as much deviation between the scores of three years after the removal of zeros and it also seems that the scores come closer to the normal distribution.

2 Introduction

In the examinations were participated students who completed having attended all the lectures in the lesson during the semester, almost all the lectures, no lecture because of the military service or because of failure of a previous examination of the same lesson. The results of the statistical analysis will be presented with statistical measures through tables and graphs for the three exam periods separately, as well as for the three whole years.

3 Data presentation

The data that will be analyzed comes from the students for the Statistics II lesson of June examinations in the periods from 1997 to 1999. Table 3.1 indicates the scores of the number of students who participated in the June examinations in the periods from1997 to1999. The writings are rated with a maximum score of 10 so that the scores range from 0 to 10. We examine the valid values of the data obtained by including the zero values, that is, the zero scores of those students who left after submitting the subjects. Specifically, in the 1999 examination, it appears that the percentage of zeros in the writings exceeds 58%, i.e. more than half of the writings were scored by zero.

Table 3.1: Frequency table by score per examination period.

Scores Examin. period	0	1	2	3	4	5	6	7	8	9	10	Total valid values	Percentage of zero values
June 1997	12	9	8	6	4	4	4	1	4	1	4	57	21,25%
June 1998	9	0	3	4	4	4	6	3	7	1	1	42	21,52%
June 1999	46	2	4	6	1	5	8	0	2	3	2	79	58,23%

4 Statistical data analysis

Table 4.1 presents the descriptive statistical measures for the three exam periods separately with and without having the zeros values.

Table 4.1 Descriptive statistical measures per examination period.

	Descriptive statistical measures (with 0)				Descriptive statistical measures (without 0)			
Years	1997	1998	1999	1997-1999	1997	1998	1999	1997-1999
Mean	3.350877	4.428571	2.126582	3.061798	4.244444	5.636364	5.090909	4.90991
Median	2	5	0	2	3	6	5	5
Standard deviation	3.131016	3.045531	3.027087	3.185703	2.932231	2.205365	2.602446	2.678363
Variance	9.803258	9.275261	9.163259	10.148702	8.59798	4.863636	6.772727	7.173628
Q1	1	2	0	0	2	4	6	3
Q3	5	7	4.5	6	6	8	9	7
Total observations	57	42	79	178	45	33	33	111

From Table 4.1 it is apparent that more students took part in the examinations of 1999 compared to the other two years, but overall the high scores were observed in 1998, with a mean of 4.4 and a median of 5, although 9 values were zeros.

Furthermore, Table 4.1 compares how different statistical measures are affected by the existence of zero value. The values of mean and median measures are significantly lower with the addition of zeros values than the expected mean and the expected median, which are 5 if we assume that our observations follow some symmetric distribution. Removing the outliers seems to improve the real picture of data and in particular the year 1999. In general, it appears that the mean and median are already above 5 and the variance decreases.

Figure 4.1 presents the results of Table 4.1 graphically using boxplots.

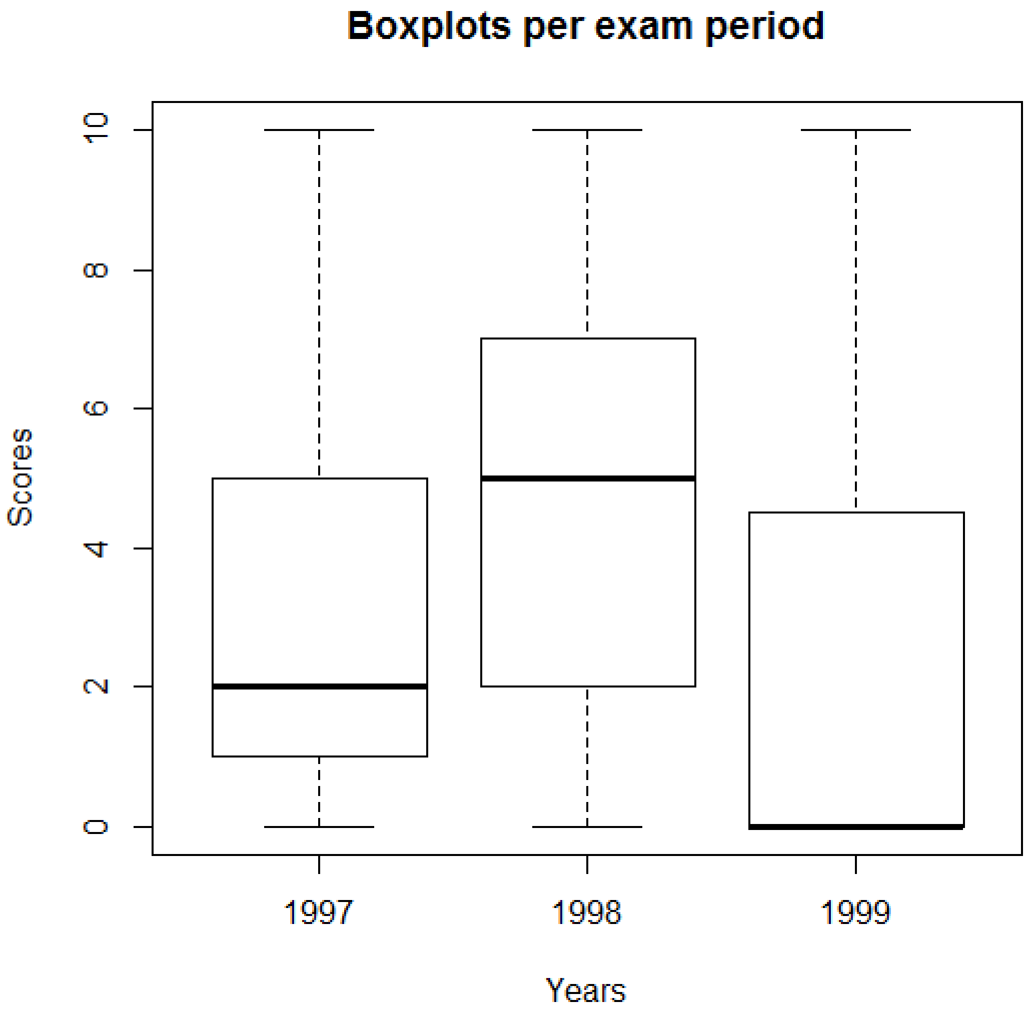

Figure 4.1: The boxplots of scores of the June exam periods from 1997 to 1999.

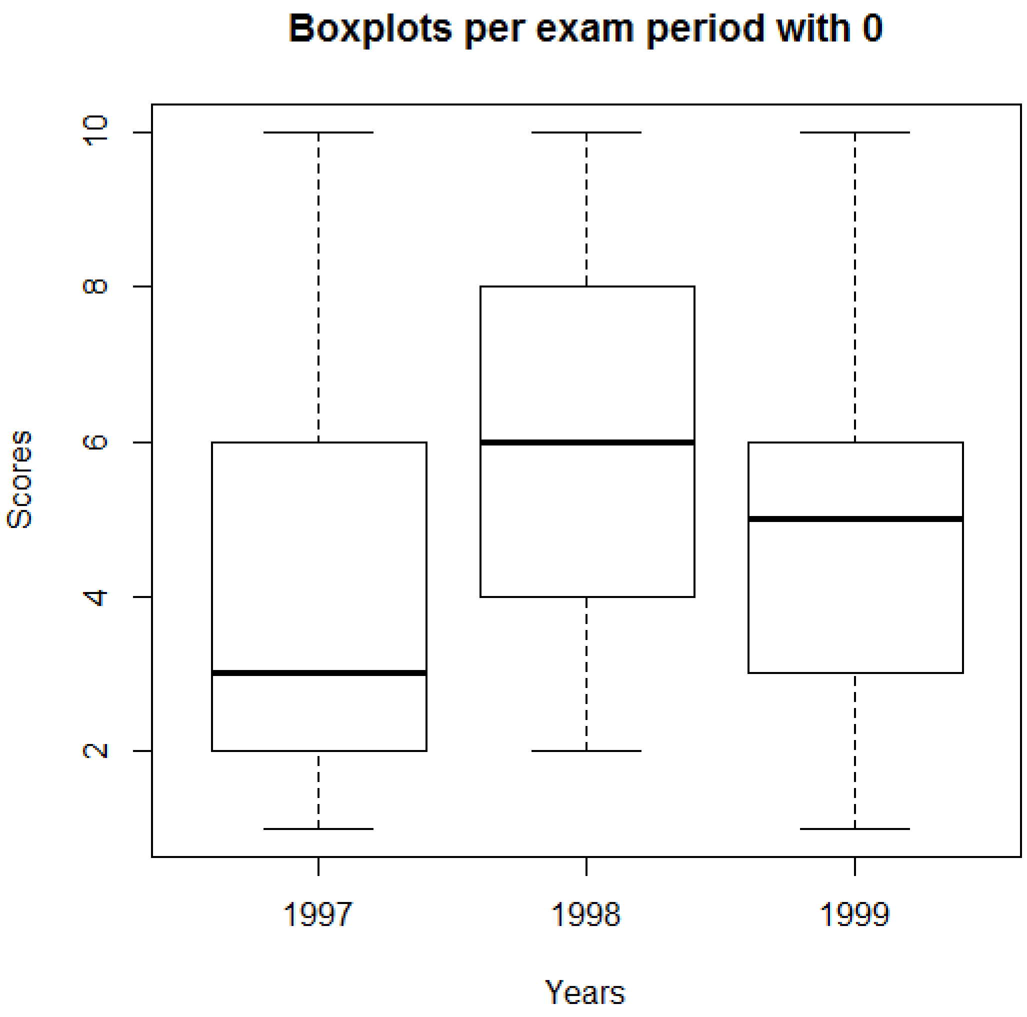

Figure 4.2: The boxplots of scores without zeros values of the June exam periods from 1997 to 1999.

Overall, comparing the data before and after the removal of zero values through the Figures or Table, it is observed that the statistical measures are more representative of the sample when we do not use the zero value. Because of the nature of our data, it is reasonable to want to exclude the zero values from our analysis as it may not truly reflect a student's ability in the course. Thus, in a future analysis, we will no longer use the zero values in our calculations.

Do our data follow normal distribution?

We start by checking our data on the assumption that we will not need the zero values and we also assume that our data follow the normal distribution. The normal distribution curve looks like a bell symmetric around the mean μ. The statistical measures to check the normality of data are presented in Table 4.2.

Table 4.2: Statistical measures for checking the normality.

Years	1997	1998	1999	1997-1999
Asymmetry	0.653582	-0.043279	0.287838	0.267759
Kurtosis	-0.796453	-1.001874	-0.855688	-0.994445
Shapiro-Wilk (p-value)	0.00035	0.128613	0.058076	7.7e-05

The asymmetry coefficient is greater than 0 (except for the 1998 period) so we claim that there is asymmetry on the right. In the 1998 year, there is asymmetry on the left. We will also need to calculate the standard asymmetry indicator for each year to check if the asymmetry affects the normality of the data. The standard asymmetry indicator for each year respectively is 1.789, 0.101, 0.675 and 1.151. The results indicate all the values less than 2, therefore the conclusion is that asymmetry does not affect the normality of the data.

In addition, we check the kurtosis indicators for each year respectively. As can be seen, their values are negative which means that we have a platykurtic distribution. The standard kurtosis indicators for each year respectively are 1.090, 1.174, 1.003 and 2.138. The first three values are below threshold 2, while the last value is above threshold 2. Thus, we conclude that kurtosis is the deviation problem from normality.

With the Shapiro-Wilk and the p-value, we reject the normality assumptions for the 1997 period only, because the value is less than 0.05. In general, we conclude from the three Shapiro-Wilk indicators that the distribution of our data is slightly different from the normal distribution. We also present graphs from the 1997-1999 periods to determine the asymmetry, the kurtosis and the degree of approximation of the normal distribution. In particular, we can observe in the Q-Q plot that some values are not on the axis so as to form an "S". This indicates that we have a platykurtic distribution. The boxplot and the histogram of the normal curve are very close to the normal distribution, that is, almost all values are around the mean.

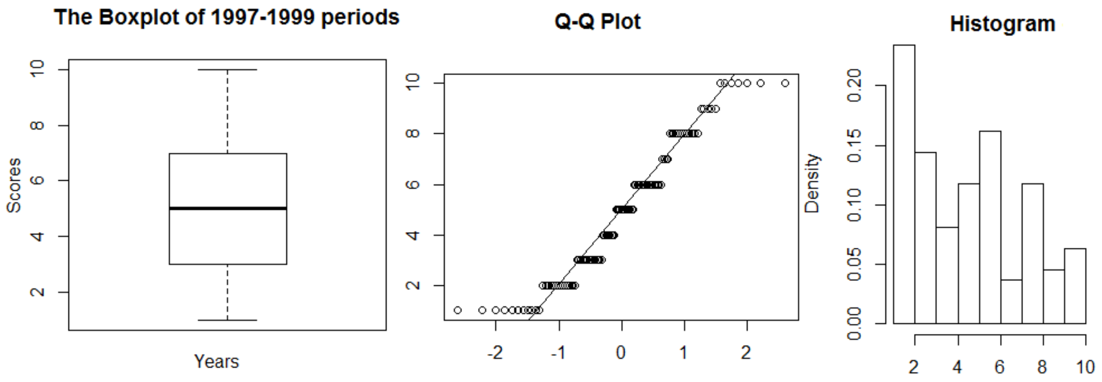

Figure 4.3: The graphs for 1997-1999 periods.

5 Conclusions

The main purpose of this study was to analyze the scores in the Statistics course for the three exam periods. Our sample had the specificity of the frequent occurrence of grade 0. Thus, by its removal, we observe that our data come closer to the normal distribution and help us draw some conclusions. The best periods in terms of performance seem to be in 1998 and 1999, in which the average performance was above 5. At the end of our analysis, we concluded that the scores in the Statistics II lesson do not follow the normal distribution.