1 Introduction

Class room sizes are a hot topic of discussion every now and then. Schools are debating about should they be investing in smaller class sizes or just keep doing what they've always done. I have a sample of data from USA's schools, which has different kinds of information about students' success and behaviour but also the size of the class that the student is in. The data is called *Star* within the package *Ecdat* in R. The data has 5748 observations (individuals) from the time period 1985-1989 in USA schools. We will be focusing only on the size of the classes and the scaled mathematical score of the student. [1]

	tmathssk	treadssk	classk	totexpk	sex	freelunk	race	schidkn
2	473	447	small.class	7	girl	no	white	63
3	536	450	small.class	21	girl	no	black	20
5	463	439	regular.with.aide	0	boy	yes	black	19
11	559	448	regular	16	boy	no	white	69
12	489	447	small.class	5	boy	yes	white	79
13	454	431	regular	8	boy	yes	white	5

Figure 1: First six observations of the dataset

Figure 1 shows the first six observations of the dataset. Variables tmathssk and treadssk represent the scaled scores of mathematics and reading tests. The classk variable represents the size of the class. A small class size is marked as *small.class*, a regular sized is *regular* and a regular sized with assistants to aid students as *regular.with.aide*. *Totexpk* shows how many years of teaching experience the teachers have. *Sex* variable shows the gender of the student and *freelunk* tells if the student is qualified for free lunch. *Race* variable tells the race of the student and *schidkn* differentiates the schools from each other.

2 Objective of this research

The objective of this research is to learn more about hypothesis testing and also to see if the class size affects the mathematical learning outcomes of students. That is, I will be comparing the measures of location of the mathematical success from two different class size students. This way we can see if the class size matters or not. Thus, I am going to use only the variables *tmathssk* and *classk* for this research.

3 Numerical descriptive statistics

This section is to tell different numerical descriptive statistics for the data used. I'm going to be referring the small sized class students as **small** and the regular sized **regular**.

	Mean	Median	Std	MAD	Min	Max
Small	491.5	489	49.47	45.96	354	626
Regular	483	478	47.90	50.41	320	626

The table above has some numerical statistics for the dataset used. Std is the standard deviation and MAD is the median absolute deviation. There are 2000 observations of regular and 1733 observations of small sized class students. Here we can see that the means are almost the same for both groups but the medians differ quite much.

4 Visual descriptive statistics

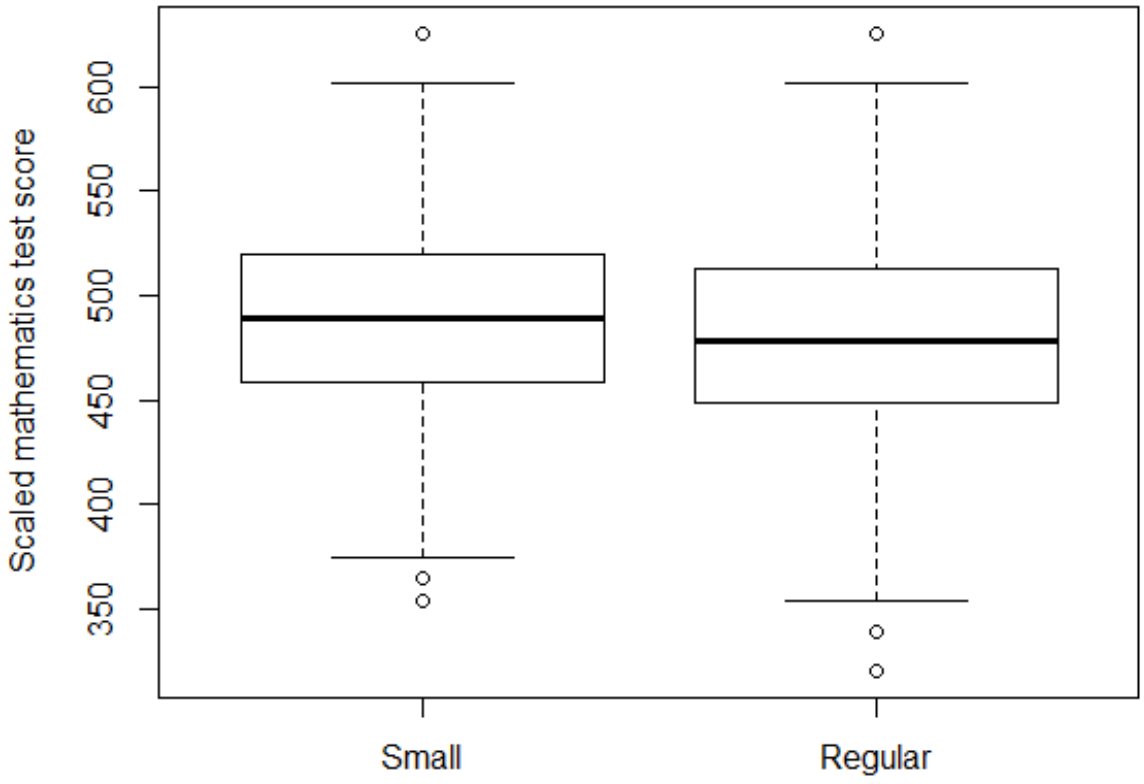

Figure 2: Boxplot for both small and regular sized classes

The boxplots in figure 2 show that the medians of the groups are really close to each other and the variances within the groups are large. There are some outliers in the groups but the number of them is rather small compared to the whole group size, thus we do not have to worry about them that much.

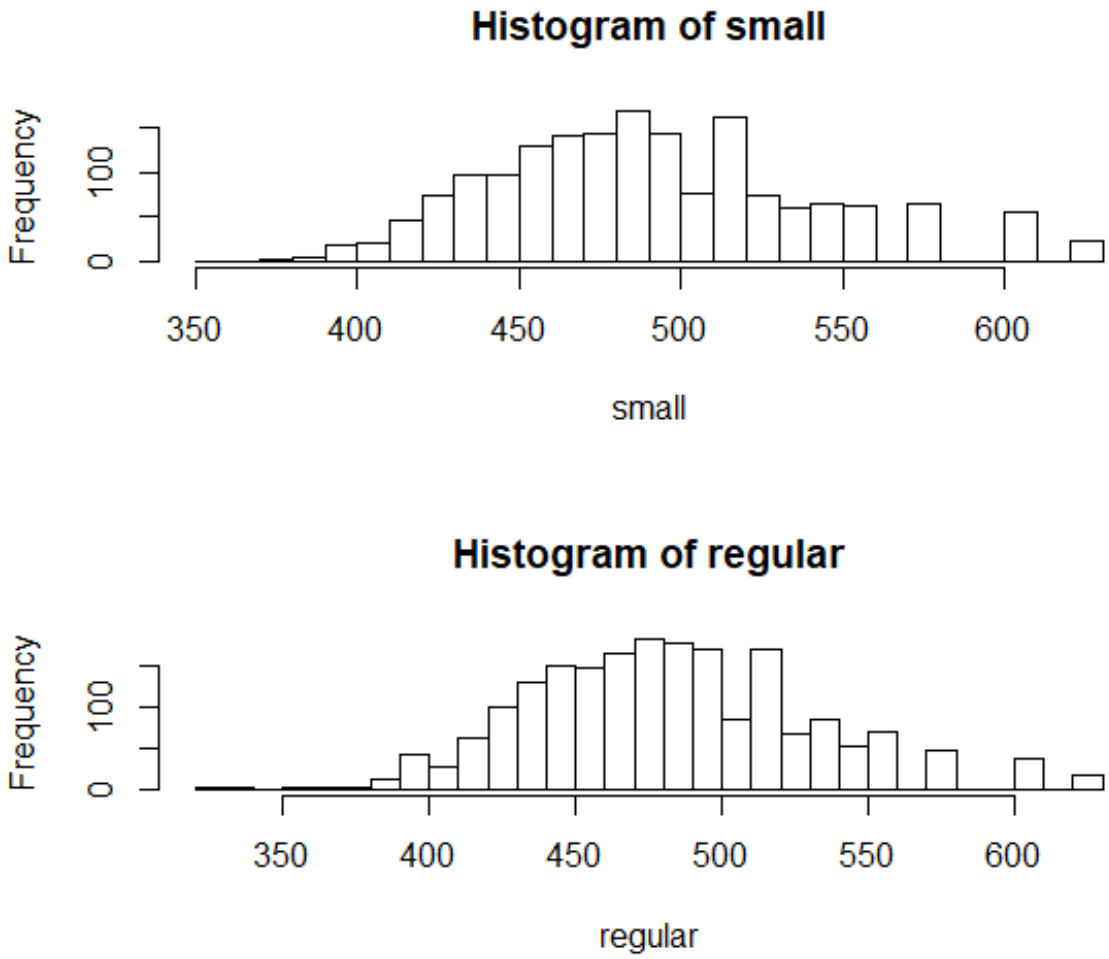

Figure 3: Histograms for both groups

Figure 3 has histograms for both of the groups. Here we can see that the distributions are almost identical up to a location shift. The histograms do not seem to follow normal distribution.

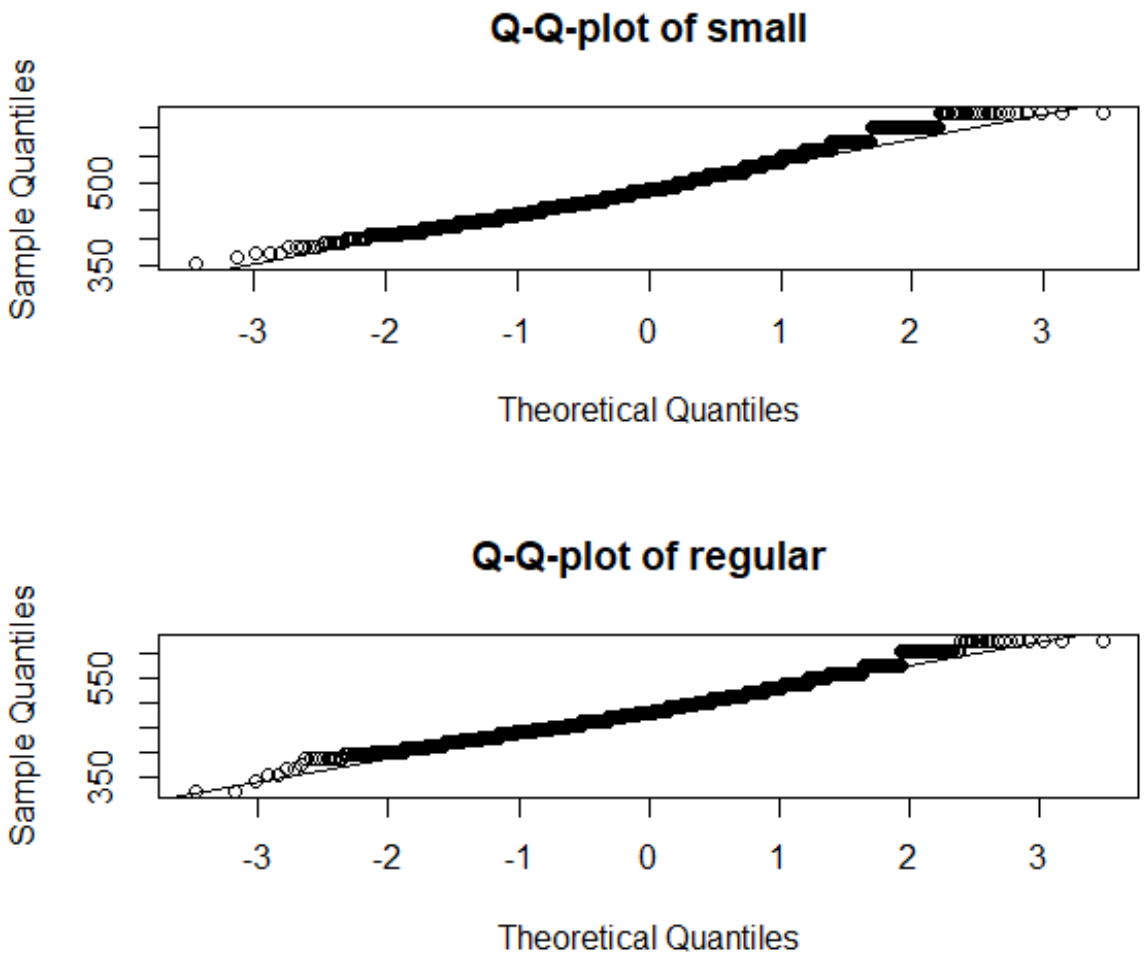

Figure 4: Quantile-Quantile -plot for both groups

Figure 4 would suggest, in contrast to figure 3, some kind of skewed normal distribution. This is due to the small image, if we zoom closer, we can see that the quantiles do not actually follow the normal-line that much in the plot.

5 Statistical method

We will be using the Two-Sample rank test also known as Wilcoxon rank-sum test to test the difference in the locations of the groups. This will reveal if the class size has effect on mathematical learning.

6 Hypotheses

The hypotheses of the Wilcoxon rank-sum test are:

$$H_0 : m_s == m_r \tag{1}$$
$$H_1 : m_s != m_r, \tag{2}$$

where m_s and m_r are the median of the variable *tmathssk* for small and regular sized classes respectively. As in, the null hypothesis is that the medians between the groups do not differ. The alternative hypothesis is that they do differ in some way.

7 Assumptions

Our Wilcoxon rank-sum test requires mild assumptions about the sample. The observations must be mutually independent and i.i.d samples from continuous distributions. The test also assumes that the distributions of the groups are equal up to a location shift. This can also be used for discrete data, which our data is but this means that it is possible that some of the sample points have the same rank. The R software will calculate the ranks so that all tied points are assigned a rank equal to the median of the corresponding ranks. [2]

8 Analysis

The significance level used in our analysis is $\alpha = 0.01$.

```
        Wilcoxon rank sum test with continuity correction

data:  small and regular
W = 1888500, p-value = 2.163e-06
alternative hypothesis: true location shift is not equal to 0
```

Figure 5: R output for Wilcoxon rank-sum test

Figure 5 shows the output for running the Wilcoxon rank-sum test for the groups small and regular. This can be interpreted as follows.

1: R software calculated the ranks correctly for discrete data
2: The data used are the correct groups, small and regular.
3: The test statistics W is 1888500 and the p-value of the test is 2.163×10^{-06}.
4: The alternative hypothesis is: true location shift is not equal to 0.

9 Checking assumptions

Now it's time to check the assumptions made previously to see that we can trust our test result. We assumed that the samples are i.i.d. This cannot be tested reasonably with any test but it can be concluded that the observations being individual students from a plethora of different schools that the samples are i.i.d. The discrete data has been dealt with by the R software. We also assumed that the distributions of the groups are equal up to a location shift. This can be seen to be true by graphically inspecting the boxplots and histograms presented previously.

We can also check that our groups are not normally distributed to see that we can't use the t-tests.

```
> jarque.bera.test(small)

        Jarque Bera Test

data:  small
X-squared = 56.846, df = 2, p-value = 4.529e-13

> jarque.bera.test(regular)

        Jarque Bera Test

data:  regular
X-squared = 53.575, df = 2, p-value = 2.325e-12
```

Figure 6: R output for Jarque-Bera test

From figure 6 we can see that the p-value for normality in both of the groups are almost zero with the Jarque-Bera test, thus they are not normally distributed.

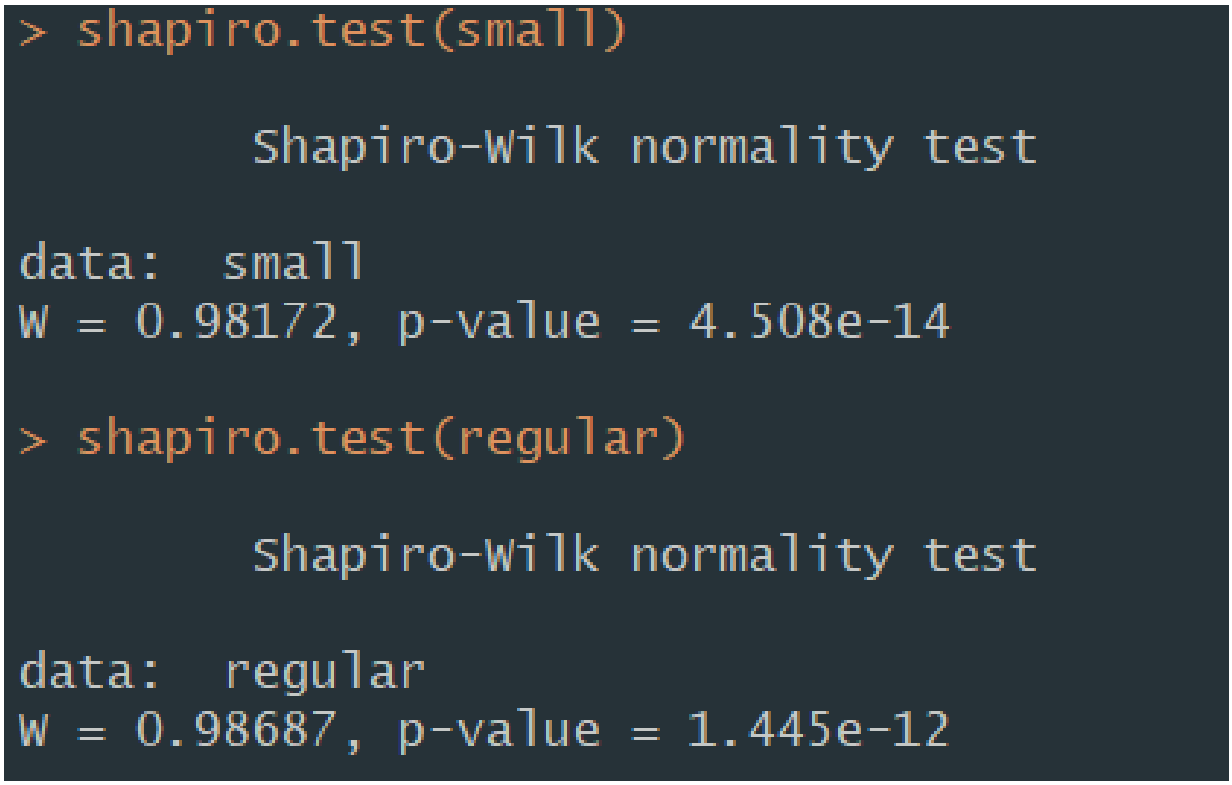

Figure 7: R output for Shapiro-Wilk test

Figure 7 shows the same result with Shapiro-Wilk test for the groups as the figure 6.

10 Conclusions mathematically

From Wilcoxon rank-sum test we got a p-value of 2.163×10^{-06}. This means that with our significance level $\alpha = 0.01$, we reject the null hypothesis that the groups have the same median value. Thus, we accept the alternative hypothesis of different medians between the groups. The type 1 error rate of our test is at most 0.01%

11 Conclusions

To conclude our research, the Wilcoxon rank-sum test gives us the result that there is a difference in the median mathematical scores of students from small and regular sized classes. The test does not give us information on the way the medians differ, only that they do differ in some way.

12 Improvements and alternative methods

There are several aspects that could improve our research about the effect of class size in mathematical learning. First of all, the dataset is outdated as the observations are from 1980s thus making them a bad sample to make any conclusions about the situation nowadays. Secondly, the data has made a distinction between the small and regular class sizes, but doesn't tell us what are the actual limits that classify a class size to be either one of the categories. By knowing the explicit border between the group sizes could make the analysis more precise. Next improvement could be that the data doesn't give any explanation how the mathematical tests are scaled. Now we must assume that they have been scaled so that every school has the same point limits for the tests. Also, the data doesn't say anything about the education level of the students, thus we do not know if the difference in the learning outcomes comes from the class size or the experience of the students. We could also improve this research by testing only one of the schools and not all of them as this would eliminate the possibility of "better" schools skewing the data.

To summarise how we could continue our analysis of the effect of class size to mathematical learning, we would have to get newer and better data. We could also focus on specific schools to get more reliable results.

References

[1] `https://rdrr.io/cran/Ecdat/man/Star.html`

[2] MS-C1620 - Statistical Inference course, Aalto-University