Assessing the Effect of Class Size on Math Scores of the First Graders

Team ID: 9

Zhikuan Quan (Model Buidling); Daidai Zhang (Interpretation and causal inference); Wenfeng Chang (Model Analysis); Jinghui Li (Exploratory Data Analysis)

Github repo:

1. Introduction

The Student/Teacher Achievement Ratio (STAR) project attracted many researchers' attention because of its large sample size and the characteristics of the study design, which was funded by the Tennessee General Assembly and conducted in the late 1980s (Word, E.R., 1990). Based on the literature review, the main characteristics of the project design are stated as below:

- The STAR project had three different level of class size as treatments that including small class (13 to 17 students per teacher), regular class (22 to 25 students per teacher), and regular-with-aide class (22 to 25 students per teacher), which started as the students entering school in kindergarten through the third grade (Achilles et al, 2008).
- There were two separate and independent randomizations happened to arrange experimental units including randomizing teachers to different class types and randomizing students to different classes/teachers (Achilles et al, 2008).
- The schools enrolled in the project had at least one class of each type for proper randomization with a sufficient number of students (Achilles et al, 2008).
- The original data file contained over 11,000 students' information and 79 schools (Achilles et al, 2008).

Under the randomization, the experiment design is unbalanced randomized complete block design. The primary question of interest is to investigate the effects of class size on teacher-level math scaled scores for the first graders.

2. Exploratory Data Analysis

The original data of 1st grade math scores were explored after removing all the missing data, resulting in 6598 observations with full information of class size, teacher ID and school ID. The summary statistics of math score are shown in Table 1. In order to make causal inference on the effect of class size on math score, Stable Unit Treatment Value Assumption (SUTVA) has to hold. Therefore, teacher (or class) was used as the unit of analysis instead of individual student, to avoid the interaction among students within one class. As shown in Figure 1, the median and mean scores for each class were very close. Mean and median are both good summary statistics in this case, but average score is more often used to evaluate a class or a teacher. Therefore, mean math scores of students taught by each teacher were calculated for further data analysis.

Table 1 Summary statistics of 1st grade math scores

Minimum	Maximum	Median	Mean	SD
404	676	529	530.53	43.11

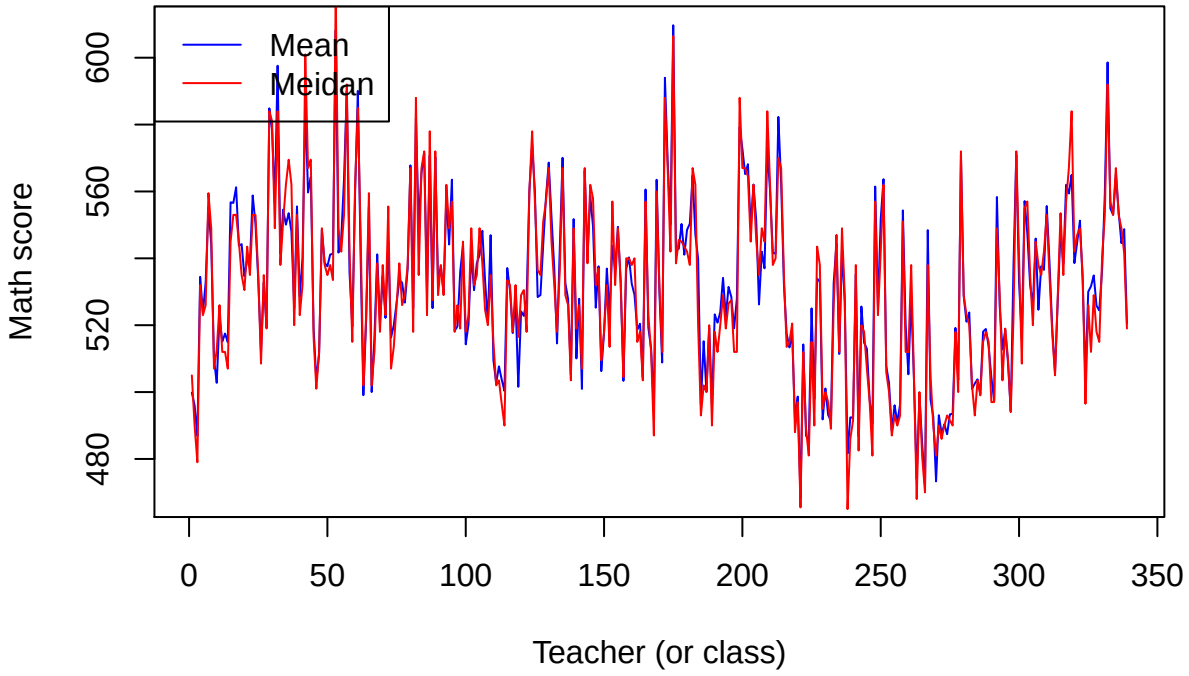

Figure 1 Means and medians of math scores for each teacher

339 mean math scores of each teacher were obtained, with 3 class sizes in 76 different schools. As shown in Figure 2, the number of math score observed for each teacher ranged from 11 to 29, with most teachers having 20 to 24 observations (45 %). The distribution of mean math score is shown in Figure 3, which is nearly symmetric. The data obtained was unbalanced as each school had different number of each class size, which might cause unequal variances in data analysis. This issue will be discussed in the following.

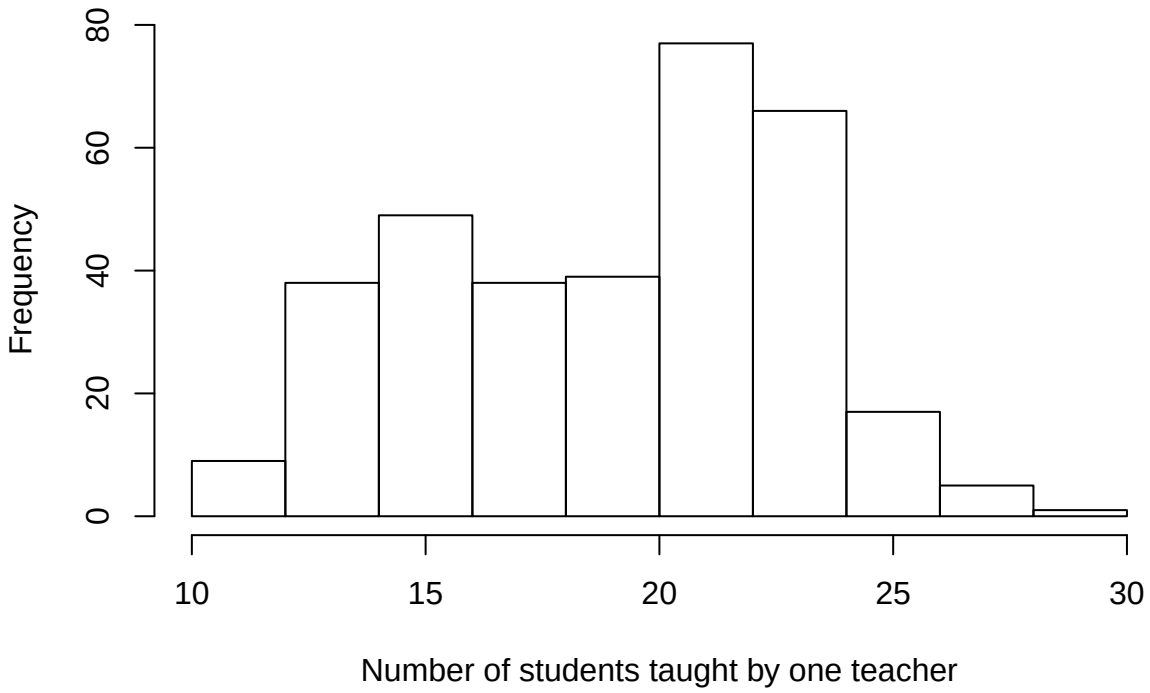

Figure 2 Histogram of the number of math score observed for each teacher

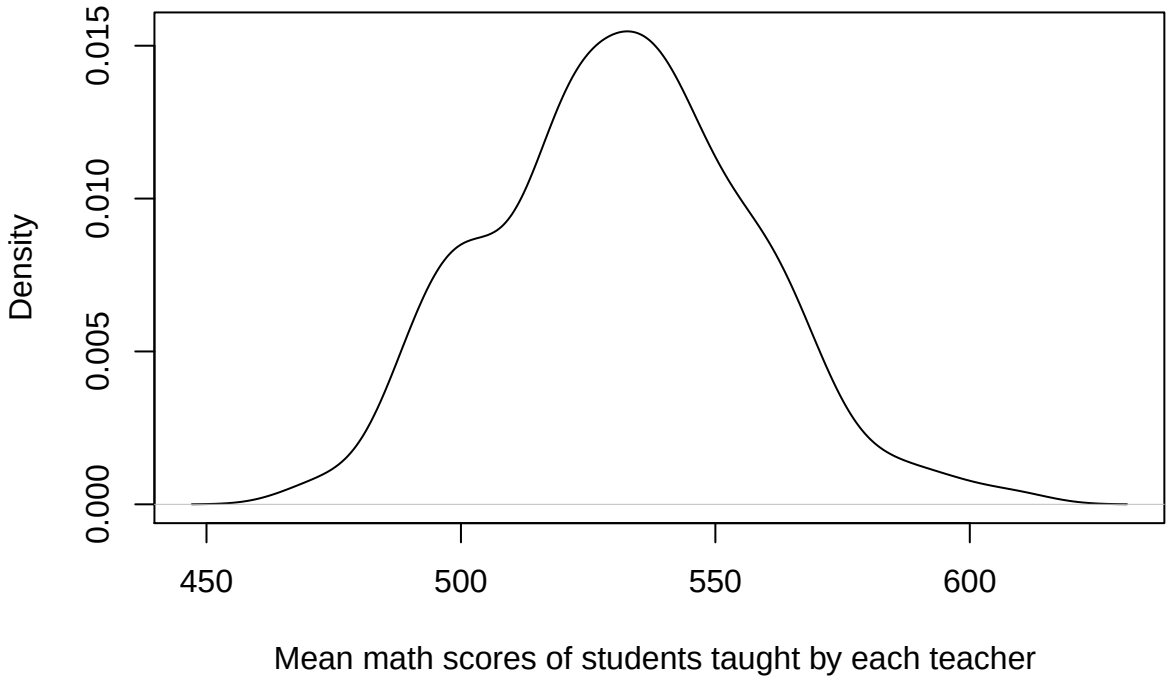

Figure 3 Distribution of mean math score of students taught by each teacher

3. Two-way ANOVA Model

3.1 Model Building

The design of experiment is a randomized complete block design, and the randomization is to assign teachers to different types of class. In order to alleviate the nuisance effect of different schools on math scores of first graders, we consider different schools (School ID) as a block factor. In this case, we mainly focus on the primary effect of class type on the teacher-level math score. We do not need to analyze the effect on one specific school and there is not enough statistical power to estimate the interaction effect of school and class type, so two-way ANOVA model without interaction terms is used to analyze. In this two-way ANOVA model, we take 3 class types and 79 schools as two factors. The school is called block factor since each school as every stratum of our experiment. The unit of analysis is the teacher or class, so we use the average math score of the students taught by each teacher as response variables. We set the model notations as below:

- Response variable Y_{ijk}: the average math scaled scores of students taught teacher k in school j of class type i;
- Class type i has 3 levels: $i = 1$ if small class; $i = 2$ if regular class; $i = 3$ if regular class with aide;
- School indicator j has 79 levels, which represents different schools.

The model equation is:

$$Y_{ijk} = \mu_0 + \mu_i + \gamma_j + \epsilon_{ijk}$$

where μ_0 is the overall average math score, μ_i is the effect due to the class type i and β_j is the effect due to the school j. In the two-way ANOVA model, we assume that the error terms ϵ_{ijk} $N(0, \sigma^2)$: (1). Normality of error terms; (2). Constant Variance of error terms; (3). Independence of error terms.

3.2 Fitting result

We use R to fit two-way ANOVA model without interaction terms. Taking small class (class type 1) as reference group, the results show that the regular class without aide and regular class with aide both tend to have lower average math score. In addition, the mean math score in regular class with aide seems to be similar or slightly higher than the regular class without aide.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	502.5	9.69	51.87	<0.05
class2	-13.37	2.19	-6.1	<0.05
class3	-11.4	2.28	-5	<0.05

3.3 Model Diagnostics

4. Statistical Inference

4.1 Scheirer–Ray–Hare Test

According to the diagnostics of two-way ANOVA model, the distribution of residuals are heavy-tailed with non-constant variance across group. In this case, we cannot use F-test or other relative normality-based statistical tests to analyze the effect of class type on math scaled scores across teachers. However, the Scheirer–Ray–Hare(SRH) test, which is an extension of the Kruskal-Wallis test, can be used to test whether the math scaled scores is affected by class types and different schools.

The SRH test is a nonparametric test which is suitable to randomized complete block design. In this case, we are not interested in the effect of class type on math scores in one specific school. Even though the conclusion and interpretation of interaction terms in SRH test is conservative, we can still use this test to examine the main effect of class type. The null hypothesis and alternative hypothesis are listed below:

H_0: All types of class have the same average math scaled scores ($\mu_1 = \mu_2 = \mu_3$).

H_a: Not all types of class have the same average math scaled scores (at least one $\mu_i \neq \mu_j$).

	Degree of Freedom	Sum of Square	H-Statistic	P-value
class	2	164742.2	17.15	0
sid	75	2042131.6	212.61	0
class:sid	146	598192.3	62.28	1
Residuals	115	441421.8	NA	NA

From the result of SRH test, given the significant level 0.05, the null hypothesis of equal sample mean is rejected significantly. It means that the type of class can affect the math scaled scores in the first grade, the effect of class type on scores is significantly different across teachers.

4.2 Multiple Comparison

Since the family-wise difference of average math scores shows significance, Dunn's Test can be used to pinpoint which specific class type tends to have higher average math score. In this case, Dunn's multiple comparison

test, which is also a nonparametric test, can analyze the pair-wise difference among different types of class. In order to maintain Type I error, Benjamini-Hochberg Procedure is applied to adjust the p-value. For each comparison, the null hypothesis and alternative hypothesis is that:

H_0: The average math scaled scores is the same between two types of class ($\mu_i = \mu_j$).

H_a: The average math scaled scores is not the same between two types of class ($\mu_i \neq \mu_j$).

Comparison	Z-Statistic	P-value	Adjusted p-value
Small Class - Regular Class	4.10	0.00	0.00
Small Class - Regular Class with aide	2.41	0.02	0.02
Regular Class - Regular Class with aide	-1.51	0.13	0.13

According to the result of Dunn's multiple comparison test, given the significant level 0.05, the smaller class tends to have higher average math scores in the first grade compared with regular class with or without aide. In addition, since the adjusted p-value for comparison of regular class with aide and regular class without aide is 0.13 (>0.05), these two types of class tends to have similar average math scores across teachers.

5. Disccusion and Conclusion