Project 2: Project STAR II

01/31/2020

1 Introduction

1.1 Background

Tennesses Student/Teacher Achievement Ratio study (Project STAR) was conducted in the late 1980s to evaluate the effect of class size on test scores. The study randomly assigned students to small classes, regular classes, and regular classes with a teacher's aide. In order to randomize properly, schools were enrolled only if they had enough study body to have at least one class of each type. Once the schools were enrolled, students were randomly assigned to the three types of classes, and one teacher was randomly assigned to one class.

The primary scientific question of interest is whether there is a treatment effect of class types on math scaled scores across teachers in 1st grade, with the school indicator as the other factor. In this study, we implement exploratory data analysis, two-way ANOVA model, model diagnostics, hypothesis testing. In the end, we conclude our findings and discuss any causal statements that could possibly be made based on our analysis and assumptions, and compare the results between Project 2 and Project 1.

1.2 Statistical questions of interest

To answer the primary scientific question of interest, we regard this project as randomized block design and propose to fit a two-way ANOVA model with the quantitative test scores grouped by teacher as the outcomes, class type as main factor and school as blocking factor. We will then run our model diagonstic to see if the assumptions of the model hold and test whether or not there is a treatment effect on the test scores.

2 Analysis Plan

2.1 Population and study design

According to the description of the dataset, over 7,000 students in 76 schools were randomly assigned into one of three different treatments: 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 with a full-time teacher's aide). Classroom teachers were also randomly assigned to the classes they would teach. We will only focus on the data about the 1st grade students as study target[1].

2.2 Descriptive Analysis

First of all, we will pick up all valuable variables we are interested in. To deal with missing value, we will draw plots to show the percentage of missing value for these variables, then decide whether drop the missing value or not. We will then explore the math scores grouped by teacher ID. In the end, we study the treatment of class type as well as school ID. Based on different class type in different school blocks, we will draw box plot on math score and check if there is any difference in different treatment.

2.3 Two-way ANOVA Model

In order to investigate whether there are effects of class types on math scaled scores in 1st grade and minimize the effects of systematic error, randomized block design is uesd, and we will create blocks by different schools. Within blocks, students and teachers are randomly assigned to class in different types. Therefore, it is possible to assess the effect from different levels of class types without worry about the influence from different schools. Although there are two factors in this experiment, we are not interested in the effect from different class type as well as different school. As a result, interaction term in this model should be excluded.

The two-way ANOVA model for a randomized block design is shown as below:

$$Y_{ijk} = \mu + \tau_i + \beta_j + \epsilon_{ijk}, \quad \text{with} \quad \epsilon_{ijk} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

By notation,

- The index i denotes main factor level. There are 3 levels about class type: `small`, `regular` and `regular+aid`. The index j denotes blocking factor level. There are 76 levels about different school ID. The index k denotes experimental unit, about mean score of k-th teachers in 1st grade.
- Y_{ijk} refers to the mean math score for k-th teacher ID in i-th type of class and j-th school ID. μ denotes the overall mean value of the mean math scores in 1st grade for all treatments. τ_i denotes main effect of level i from the class type factor. βj denotes blocking effect of level j from the school ID factor. ϵ_{ijk} denotes random errors.

Moreover, there is a "sum-to-zero" constraint: $\sum_i \tau_i = \sum_j \beta_j = 0$.

Assumptions of two-way ANOVA model specified as follows:

- The random errors are assumed to be identically and independently distributed from a normal distribution with mean 0 and variance σ^2.

2.4 Model Diagnostics

This part will check the assumptions of our one-way ANOVA model so that this model is appropriate. To verify the independence assumption, we would analyze the experiment design to check the randomization of the samples. For the residuals, we would use the residuals v.s. fitted values plot to check equal variance assumption. Then, Q-Q plot is drawn to check whether the residuals are normally distributed or not.

2.5 Hypothesis Testing

In this part, in order to investigate whether there is a treatment effect based on class types, we decide to use F-test. Moreover, if there is a treatment effect, we will investigate comparisons between two different class types. Because our dataset is unbalanced for each group, Tukey's Procedure is the best method to test among Tukey, Bonferroni and Scheffe. Moreover, Tukey's Procedure has the smallest T-statistics, which means it has the narrowest confidence interval, then it can give a more precise estimation of the difference.

3 Results

3.1 Descriptive Analysis

The dataset we investigate has 11601 rows and 379 columns, and we choose four variables: `g1classtype`, `g1schid`, `g1tchid`, `g1tmathss` which are related with math scaled scores in the 1st grade. `g1classtype` represents the class type of students in 1st grade. `g1schid` represents school ID. `g1tchid` represents teacher

ID. `g1tmathss` represents math scaled scores of students in 1st grade. See Appendix III for more figures and detailed explanation.

3.2 Two-way ANOVA Model

Table 1 shows a summary of our ANOVA model. The mean of square among class types is 5809 while that among school is 1824. As is known that the larger the mean of square is, the larger the variability among the group is. Therefore, the variability of mean math score in different class types is much larger than that in different school. In addition, the mean of square of residuals is 277, which is much smaller compared with class types and school. And the p-values of the two variables are much smaller than 0.001, which are highly significant in statistics.

Table1: two-way ANOVA table

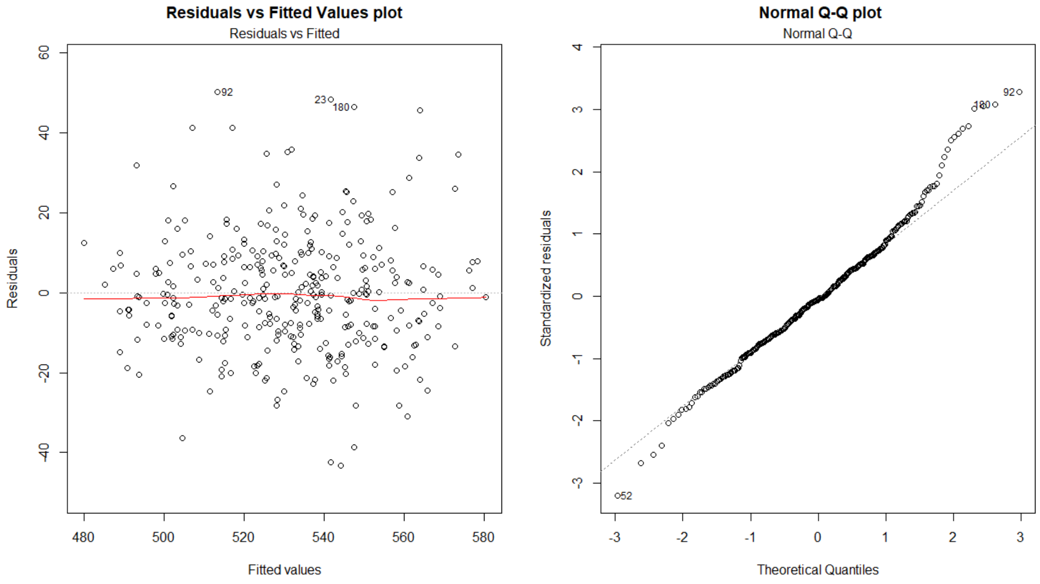
Source of Variation	Degree of Freedom	Sum Square	Mean Square	F value	P value
g1classtype	2	11617	5809	20.99	3.52×10^{-9}
g1schid	75	136833	1824	6.59	$< 2 \times 10^{-16}$
Residuals	261	72225	277		

3.3 Model Diagnostics

In this experiment, researchers assigned students to small classes, regular classed and regular classes with aide randomly. Besides, teachers are also randomly assigned to classed in each school. By adding the randomized block design of schools in the experiment, there are adequate reasons to believe that the outcomes do not depend on each other. In other words, the model satisfies the independence assumption.

From the Residuals vs Fitted Values scatter point, these points are uniformly distributed on both sides of x-axis, which means that our model does not violate the equal variance assumption. The Normal Q-Q plot illustrates that the residuals are slightly heavy-tailed, but given the large sample size, we can verify that our model meet the normality assumption.

Figure 1: residuals vs fitted value and Q-Q plot

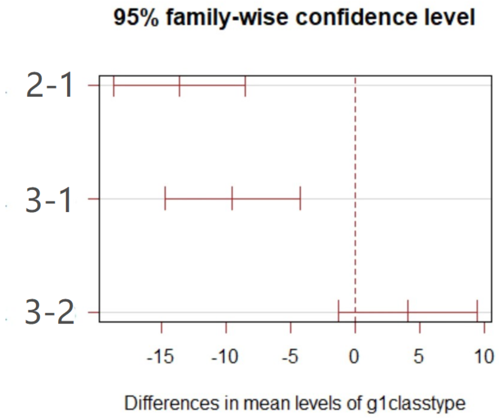


3.4 Hypothesis Testing

For F-test part, to test null hypothesis $H_0 : \tau_1 = \tau_2 = \tau_3$ against alternative hypothesis H_a: not all τ_i are equal. We compute F-statistic: $F^* = \frac{MSTR}{MSE} = 21$ and $F(0.95, 2, 6597) = 3$. Since $F^* > F(0.95, 2, 6597)$, we can reject the null hypothesis at the significance level 0.05. We can claim that there exists a treatment effect based on class types.

For Tukey's Procedure part, we use Tukey's Procedure to test difference between each group on class types. From figure 2, we can get the confidence intervals for all three pairs of class types.

Figure 2: the plot of 95% family-wise confidence intervals of the difference all three pair of class types by Tukey's Procedure



As we can see, 0 is in the 95% family-wise confidence interval of the difference between regular aide class and regular class. It means that the difference between these two groups is really small and it is close to 0. From table 2, we can find that regular class and small class have the largest difference. Teachers who are teaching in small class have 13.58 scores higher than regular class on the average. And regular aide class and regular class have the smallest difference.

Table 2: the table of 95% family-wise confidence intervals of the difference all three pair of class types by Tukey's Procedure

Pairwise Comparison	Difference between two means	Lower Bound	Upper Bound	Adjusted P value
Regular vs Small	-13.58	-18.66	-8.50	3.65×10^{-9}
Regular + Aide vs Small	-9.49	-14.76	-4.22	9.09×10^{-5}
Regular + Aide vs Regular	4.09	-1.27	9.46	0.17

3.5 Causal Analysis

In this part, we use average treatment effect to measure the causal effect. We can make causal statements if it follows two assumptions:

Stable unit treatment value assumption: The potential outcomes for any unit do not vary with the treatments assigned to other units, and for each unit, there are no different forms or versions of each treatments level, which leads to different potential outcomes.

Ignorability: Treatment assignment is independent of the outcomes.

For the first assumption, it means that we could exclude the probability that a teacher is assigned or not assigned to small class does not affect the mean of math grades in other class types. Also, same version of treatment needs to be ensured, which means teachers in the class of the one specific type give consistent education of the class type. From the background of project STAR, we know that it is a randomized experiment since it randomly assigns students to different class types and teachers are also randomly assigned to class types. Therefore, both of two assumptions hold based on the setting of the experiment. However, for the second assumption, treatment assignment is not independent of the outcomes. Since project STAR is a randomized block design but not a completely randomized design, students and teachers are only randomized in each block, and they are not randomly assigned to schools. Therefore, treatment assignment and the outcomes are not independent. We cannot make any causal statements due to violating ignorability assumption.

4 Discussion

4.1 Difference between two projects

In project 1, we assume the experiment to be completely randomized design. That is, only class type itself effects the math scores in 1st grade. Other factors like gender and lunch which are randomized in this experiment do not affect treatment. And then, we use one-way ANOVA model to investigate it. In project 2, we regard this experiment as randomized block design and use school as block, which is closer to the real world. We take account of schools' effects and build a two-way ANOVA model to investigate the class types' effects on math scores.

In terms of the results, in project 1, we obtain that all of pairwise comparisons of math scores in three class types are statistically significant difference. In project 2, we cannot make a claim that there is a statistically significant difference between the math scores in regular class and regular with aid class. The difference may be caused by different statistics served as the value of outcome in experiment setting. More important, in project 2, blocking strategy can mitigate the effects of systematic error. From the data exploration, different schools also affect math scores due to educational resources differences. Within blocks, it is possible to assess the effect of different levels of the class types without having to worry about variations due to different schools.

Moreover, for causal inference, we make some causal statements in project 1, but we fail to make any causal statements in project 2. Because when using average treatment effect to measure the causal effect, Ignorability assumption are only satisfied in completely randomized experiment. In randomized block design, treatment assignment is no longer independent of the outcomes.

4.2 Equal Variance

In model diagnostics, we verify the equal variance assumption by observing residual plot. Because nearly half of cells only have one observations in the model, using Levene's test, Hartley's test or Bartlett's test to check equal variance does not make much sense. Since we do not know other feasible tests, we leave it for further study.

4.3 Causal Inference

In causal inference, we use average treatment effect to measure the causal effect. Because of violating ignorability assumption, we cannot make any causal statements. However, there are many other ways to measure the causal effect with other assumptions. They can be feasible for incompletely randomized design. As far as we concerned, we fail to get the causal statement.

5 Appendix I. Reference

[1] C.M. Achilles; Helen Pate Bain; Fred Bellott; Jayne Boyd-Zaharias; Jeremy Finn; John Folger; John Johnston; Elizabeth Word, 2008, “Tennessee’s Student Teacher Achievement Ratio (STAR) project”, https://doi.org/10.7910/DVN/SIWH9F, Harvard Dataverse, V1, UNF:3:Ji2Q+9HCCZAbw3csOdMNdA== [fileUNF]

6 Appendix II. Session information

```
## R version 3.6.2 (2019-12-12)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 17763)
## 
## Matrix products: default
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## loaded via a namespace (and not attached):
##  [1] compiler_3.6.2  magrittr_1.5    tools_3.6.2     htmltools_0.4.0
##  [5] yaml_2.2.0      Rcpp_1.0.3      stringi_1.4.3   rmarkdown_2.0  
##  [9] knitr_1.26      stringr_1.4.0   xfun_0.11       digest_0.6.23  
## [13] rlang_0.4.2     evaluate_0.14
```

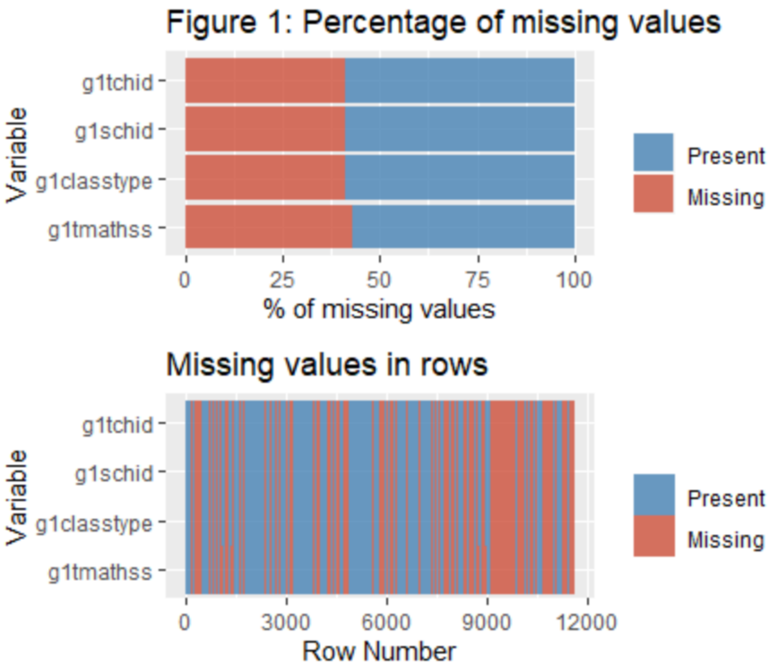
7 Appendix III. Descriptive Analysis

The dataset we investigate has 11601 rows and 379 cloumns, and we narrow it down to four columns, that is, g1classtype, g1schid, g1tchid, g1tmathss which are related with math scaled scores in the 1st grade. g1classtype represents the class type of students in 1st grade. g1schid represents school’s id. g1tchid represents teacher’s id. g1tmathss represents math scaled scores of students in 1st grade. Table 1 shows the basic information of four variables.

Table 1: basic information of variables

Name	Type	The.number.of.levels
g1classtype	qualitative variable	3
g1schid	qualitative variable	76
g1tchid	qualitative variable	339
g1tmathss	quantitive variable	NA

In Figure 1, the above plot shows the percentage of missing value of four variables are nearly same, and the plot below shows that the location of missing values for three qualitative variables are consistent. It can be explained that because of low grades in kindergarten in STAR project, students quit STAR project and go to another school, or they join STAR project after 1st grade. Because we don’t know how to tackle missing values, considering that we have enough data and believe that it won’t have the significant impact on the results, we drop them directly.



In this project, we are interested in math scaled scores in the 1st grade with teachers as the unit. After verifying each teacher are only employed in specific school and teach one specific class type, we select the mean of math scaled scores of one particular class to judge teachers. Therefore, we have a new dataset which has 339 rows and four columns including g1classtype, g1schid, g1tchid and mean. Figure 2 shows the distribution of the mean of math score by different teacher. It is bell-shaped and roughly closed to normal distribution.

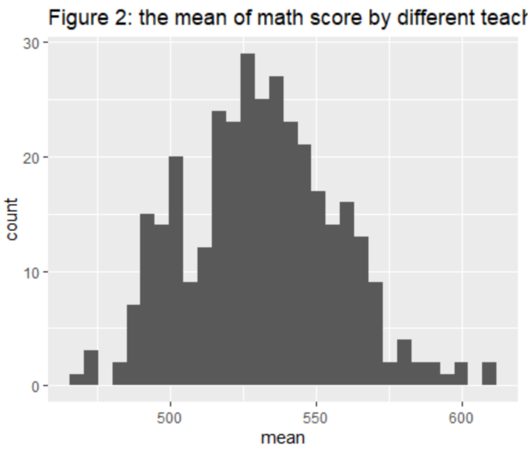


Figure 3 shows the frequency number of class type and school id. Due to too many levels in g1schid, x axis labels in the plot of g1schid is omitted for readability. We can find that g1classtype data is roughly balanced, and there is one specific school whose frequency number which is much higher than rest of schools. Figure 4 shows that the frequency number of class of different types in each school. Excepts a few extremely cases, data is more balanced than data group by class type.

Figure 3: The frequency bar plots of class type

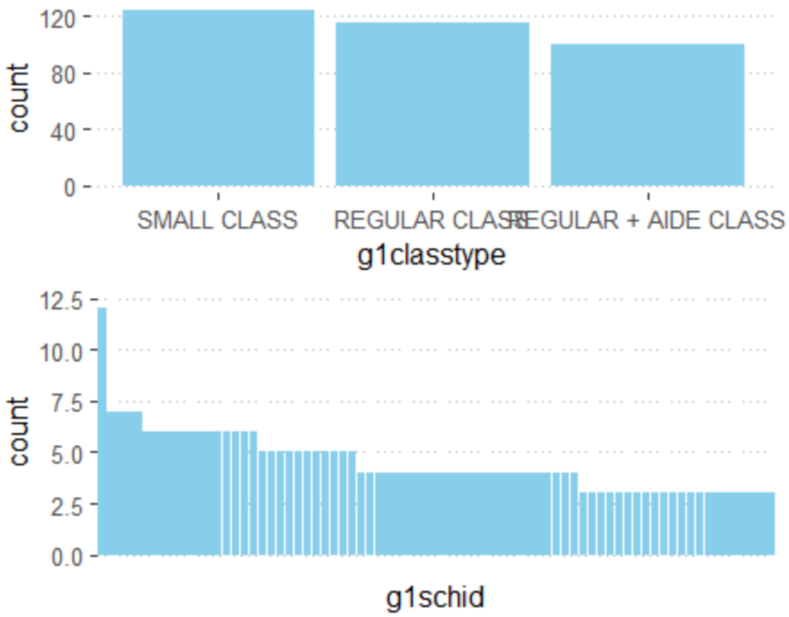
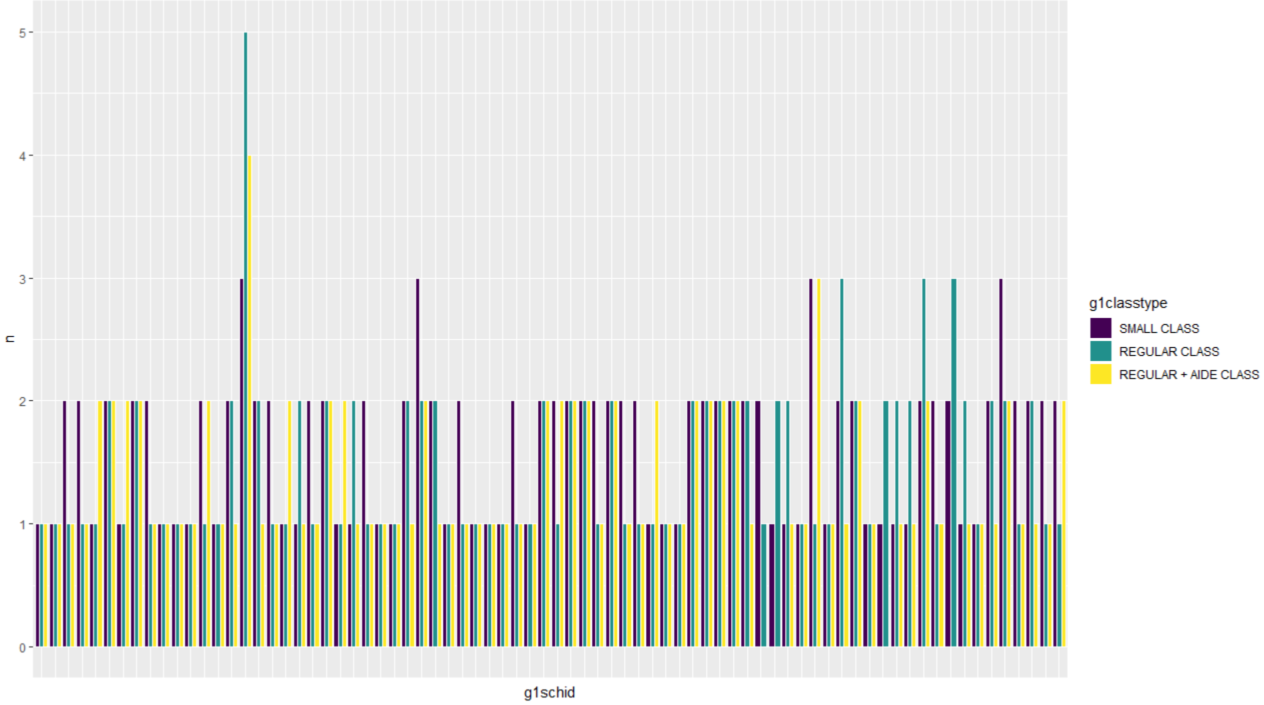
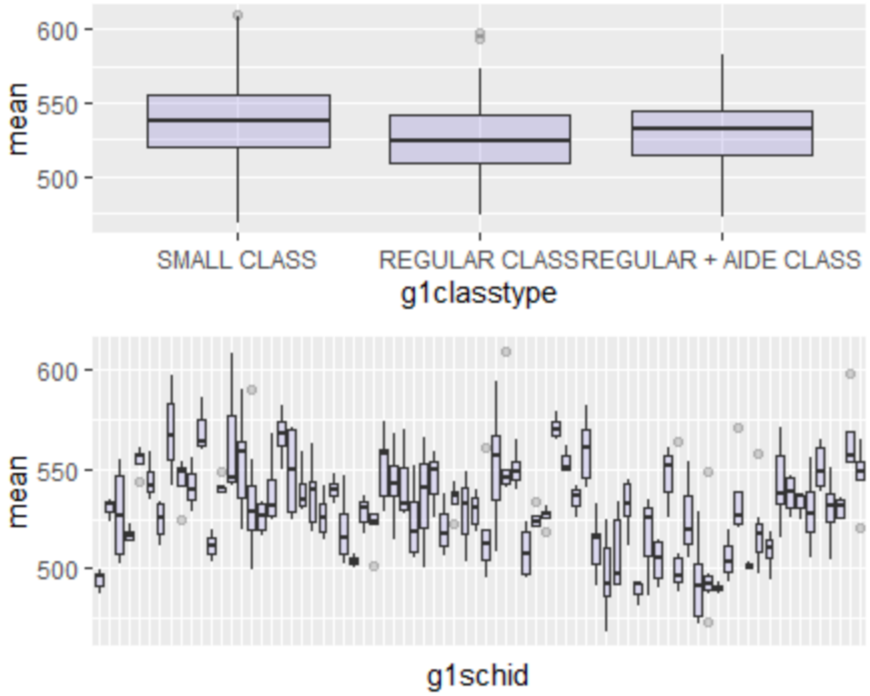


Figure 4: The frequency bar plots of grouped contingency tables



In Figure 5, it can be seen that different class types have effects on mean scores of math. Therein, teachers in small class have the higher mean scores of math and teachers in regular class have the lower mean scores of math in general. Also, different schools have impacts on mean scores of math. Aming to investigate the class types effect on mean scores of math, it is a good strategy that using schools to create blocks for eliminating effects between different schools.

Figure 5: The distribution of mean scores of math


mean
SMALL CLASS
REGULAR CLASS
REGULAR + AIDE CLASS
g1classtype
mean
g1schid