Love Matters: Understanding the influence of romantic relationships on college students' mathematical achievement

Course: STAT 344 101
Instructor's Name: Lang Wu
Date: November 12, 2020

Member	Contribution
Capri Kong (59181164)	background research, writing, data analysis
Qian Xu (95762316)	writing
Yining Guo (23418015)	data analysis
Kaili Tu (34651208)	data analysis
Group Leader:	Contribution
Rebecca Huang (94353935)	data analysis

1 Introduction

Extensive research sought to explain students' performance in STEM (science, technology, engineering, and mathematics) fields. Much of it has considered factors including gender, socioeconomic status (SES), race and ethnicity (Troncoso et al., 2016; Riegle-Crumb & Grodsky, 2010). For instance, children from low-income families are positively correlated to poor mathematical outcomes than children from middle-income families, who generally engage in more meaningful and learning extra-curricular activities (Jordan et al., 2009). Emerging research in Social Psychology also started investigating children's interpersonal relationship and its potential impacts on academic performance (Brown & Bakken, 2011; Jeynes, 2005). Social psychology studies the association between persons, amounting to awareness of one's social identities, cultivating abilities and behaviours to consolidate their relations to society. For instance, friendships with academically-oriented peers are associated with a decrease in behavioural problems and an increase in school performance (Crosnoe & Elder, 2003). While parent and peer influences on academic achievement are well documented, little research has examined links to romantic involvement during the adolescence. The limited literature has directed most of the attention towards the negative outcomes of adolescent romantic relationship involvement, such as increases in depression symptoms (Quatman et al., 2001). Yet, the role of romantic involvement on academic achievement has not been systematically investigated.

The study objective is to fill in this limitation in previous work and examine if romantic involvement has an impact on adolescents' academic achievement. The research question is how many students with romantic involvement have a passing mathematical performance. The study is of policy-making and research interest, to better understand what contributed to STEM educational gaps among students of diverse social relations and how social services and youth programs can better assist youth in navigating relationships and academic learning.

2 Data Collection

The study uses online dataset Social, gender and study data from secondary school students from UCI Machine Learning (2016) to assess the relationships between romantic involvement and mathematical achievement. The data were obtained in a survey of 395 students enrolled in a maths course at Gabriel Pereira or Mousinho da Silveira secondary schools, and was downloaded into a CSV file on Nov 11, 2020, via Kaggle.

In this study, the targeted population is all college students aged 15-22 registered in a math course at Gabriel Pereira college and Mousinho da Silveira college in 2008. The population size is N = 395, and the sample unit is individual (student). We are interested in two parameters a) average b) proportion, and to study two different types of populations 1) average math score (continuous) 2) proportion of students with a passing grade (math score ≥ 10) (binary),

based on the same samples.

Variable	Description	Type	Levels
G3	Students' math score	Numerical	[0, 20]
romantic	Students with a romantic relationship	Categorical	Yes/No

For our study, we set a research confidence level to 95% and thus a Z-score (z) at 1.96. We also set our population proportion to worse-case p = 0.5 and a Margin of Error (MOE) at 5%. Based on our design, here we obtain our sample size (n):
Sample size based on infinite population:

$$\begin{aligned} n_s &= (z^2 * p(1-p))/MOE^2 \\ &= (1.96^2 * 0.5(0.5)/0.05^2 \\ &= 384.16 \end{aligned}$$

Adjusted sample size based on required population size N=395, to the nearest integer:

$$\begin{aligned} n &= n_s/[1 + (n_s - 1)/N] \\ &= 384.16/[1 + (384.16 - 1)/395] \\ &= 195 \end{aligned}$$

Based on the calculation above, our necessary sample size is n = 195.

3 Simple Random Sampling

The first sampling method applied in this study is the Simple Random Sampling (SRS). By definition, a simple random sample is a subset of a statistical population in which each member of the subset has an equal probability of being selected as the actual sample, without replacement from the finite population. This satisfies the IID assumption, for independent, of identically distributed random variables.

Simple random sampling has clear advantages in research and experimental studies. Given a fulfilled IID assumption, a simple random sample is meant to be an unbiased representation of a group. SRS is also desirable as it often results in samples that are representative of the population, from which they were drawn on all imaginable variables. In practice, SRS is generally used in conducting randomized control tests or blinded experiments. Yet, SRS requires cautious implementation. A sampling error can occur with a simple random sample if the sample does not accurately reflect the population. When a simple random sample of the population lacks inclusivity of population elements, the representation can be skewed.

For this study, we implement a simple random sampling method to our data:

$$N = \text{Total number of units in population}$$
$$n = \text{Number of necessary sample size}$$
$$N = 395$$
$$n = 195$$
$$\text{Randomly select 195 samples} = \binom{395}{195}$$

4 Stratified Random Sampling

The second sampling method applied in this study is the Stratified Random Sampling. Stratified Random Sampling is a sampling design in which prior information regarding the population is used to determine disjoint subgroups, or strata. It is used to estimate population parameters efficiently when there is substantial variability between sub-populations.

The strata are to be allocated according to an auxiliary variable that is correlated with the variable to be estimated. In addition, relatively homogeneous group of elements within each stratum should be provided.

We choose "the presence of romantic involvement" as the auxiliary variable based on the assumption that romantic relationship should be highly correlated with the mathematical performance being measured.

$$H = \text{number of strata} = 2$$
$$N = \text{total number of units in population} = \sum_{h=1}^{H} N_h = 395$$
$$N_h = \text{total number of units in stratum h} \rightarrow N_{romantic} = 132, N_{nonromantic} = 263$$
$$y_{ij} = \text{student being selected and has a 10 or above math score}$$
$$n_{romantic} = \text{weighted sample size for students with romantic involvement}$$

Therefore, $\frac{N_h}{N} = \frac{n_{romantic}}{n} \rightarrow \frac{132}{195} = \frac{n_{romantic}}{132} \rightarrow n_{romantic} = 65$

Given $\text{n}_{romantic} = 65, n_{nonromantic} = 195 - 65 = 130$

Randomly select 65 $\text{n}_{romantic} samples = \binom{132}{65}$

Randomly select 130 $\text{n}_{nonromantic} samples = \binom{265}{130}$

5 Data Analysis

5.1 Estimation Results

We use vanilla estimation to estimate the population parameters (mean, proportion). We will apply FPC due to $\frac{n}{N} = \frac{195}{395} \geq 0.05$. For interpretation purposes,

results are corrected to 4 decimal places.

5.1.1 Estimating population mean

$\bar{y}_{\text{vanilla}} = \sum_{n=1}^{195} y_i/n = 10.4821$
$se(\bar{y}_{vanilla}) = \sqrt{(1 - \frac{n}{N})\frac{S_s^2}{n}} = 0.2449$
95% Confidence Interval: $\bar{y}_{\text{vanilla}} \pm z_{0.975} \times se(\bar{y}_{vanilla}) = [10.0021, 10.9620]$
$MOE = z_{0.975} \times se(\bar{y}_{vanilla}) = 1.96 \times 0.2449 = 0.4800$

Based on simple random sample data, we estimate that the population mean is 10.4821. Given a 95% confidence level and a standard error at 0.2449, the margin of error of the vanilla estimate is 0.4800; and the 95% confidence interval is 10.0021 to 10.9620. That is, we assume our 95% confidence interval [10.0021, 10.9620] would contain the population mean over repeated random sampling.

5.1.2 Estimating population proportion

y_{ij} = student being selected and has a 10 or above math score

$\hat{p}_s = \sum_{n=1}^{195} y_{ij}/n = 0.6718$
$se(\hat{p}_s) = \sqrt{(1 - \frac{n}{N})\frac{\hat{p} \times (1-\hat{p})}{n}} = 0.0239$
95% Confidence Interval: $\hat{p}_s \pm z_{0.975} \times se(\hat{p}_s) = [0.6249, 0.7187]$
$MOE = z_{0.975} \times se(\hat{p}_s) = 1.96 \times 0.0239 = 0.0468$

Based on simple random sample data, we estimate that the population proportion is 0.6718 . Given a 95% confidence level and a standard error at 0.0239, the margin of error of the vanilla estimate is 0.0468; and the 95% confidence interval is 0.6249 to 0.7187. That is, we assume our 95% confidence interval [0.6249, 0.7187] would contain the population proportion over repeated random sampling.

5.1.3 Estimating population mean

$\bar{y}_{\text{str}} = \sum_{h=1}^{2} \frac{N_h}{N} \times \bar{y}_{sh} = 10.5576$
$se_{\text{str}} = \sqrt{(1 - \frac{n_h}{N_h})\frac{S_{sh}^2}{n_h}(\frac{N_h}{N})^2} = 0.2232$
95% Confidence Interval: $\bar{y}_{\text{str}} \pm z_{0.975} \times se_{\text{str}} = [10.1202, 10.9950]$
MOE = $z_{0.975} \times se_{\text{str}} = 1.96 \times 0.2232 = 0.4375$

Based on stratification data, we get that mean of stratified sampling is 10.5576. Given a 95% confidence level and a standard error of stratified sampling at 0.2232, the margin of error of the stratified estimate is 0.4375; and the 95% confidence interval is 10.1202 to 10.9950. That is, we assume our 95% confidence interval [10.1202, 10.9950] would contain the population mean over stratification.

5.1.4 Estimating population proportion

$\hat{p}_{\text{str}} = \sum_{h=1}^{2} \frac{N_h}{N} \times \hat{p}_{sh} = 0.6709$
$se_{\text{str}} = \sqrt{(1 - \frac{n_h}{N_h}) \frac{\hat{p}_{sh} \times (1 - \hat{p}_{sh})}{n_h} (\frac{N_h}{N})^2} = 0.0238$
95% Confidence Interval: $\hat{p}_{\text{str}} \pm z_{0.975} \times se_{\text{str}} = [0.6242, 0.7176]$
MOE = $z_{0.975} \times se_{\text{str}} = 1.96 \times 0.0238 = 0.0466$

Based on stratification data, we estimate that the proportion of stratified sampling is 0.6709. Given a 95% confidence level and a standard error of stratified sampling at 0.0238, the margin of error of the stratified estimate is 0.0466; and the 95% confidence interval is 0.6242 to 0.7176. That is, we assume our 95% confidence interval [0.6242, 0.7176] would contain the population proportion over stratification.

5.2 Estimate Evaluation

The section evaluates the estimation results from both Simple Random Sampling Method and Stratified Random Sampling Method, and compare their strengths and weaknesses in estimation and sampling strategy. We are able to obtain the true values of the population mean and proportion, given a dataset of the targeted population N = 395.

Population	Mean	Proportion
True value	10.4152	0.6709

The table below summarises the estimated population mean and population proportion, the corresponding standard error, 95 % confidence interval and width. Based on the results, the SRS estimated population mean is closer to the true population mean than STR; however STR has a smaller value of SE and narrower CI width than the SRS estimates. Moreover, the STR estimated population proportion is exactly the same to the true population proportion, and has a slightly smaller value of SE and narrower CI width than SRS estimates.

Estimation	Mean	SE	CI.lower	CI.upper	CI.width
SRS	10.4821	0.2449	10.0021	10.962	0.9599
STR	10.5576	0.2232	10.1202	10.995	0.8748
Estimation	Proportion	SE	CI.lower	CI.upper	CI.width
SRS	0.6718	0.0239	0.6249	0.7187	0.0938
STR	0.6709	0.0238	0.6242	0.7176	0.0934

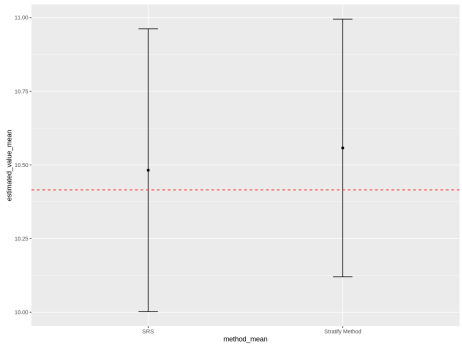

Figure 1: Estimated mean using SRS and STR and their 95% C.I.

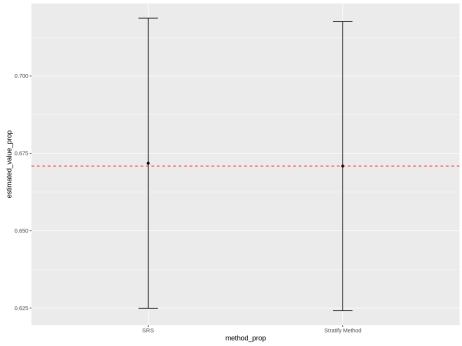

Figure 2: Estimated proportion SRS and STR and their 95% C.I.

Our results show that stratified random sample estimation offers a smaller value of standard error and a more precise confidence interval with a narrower confidence interval width, compared to simple random sample estimation. To sum up, the randomisation in stratified random sample achieve an unbiased sample, and stratified estimation is relatively efficient with a smaller variance (or standard error) than the SRS vanilla estimation in the study, which gives a more precise and better estimation. Besides, such evidence supports that students' romantic involvement is an auxiliary variable and is a considerable indicator in estimating students' average math score and proportion with passing grades.

There are significant advantages and disadvantages to applying Simple random sampling and Stratified random sampling in our study. Simple random sample advantages include the ease of use and accuracy of representation. SRS requires no division of the population into sub-populations, rather only require careful governance of the randomness in the selection process, to ensure each unit of the population has an equal probability of being chosen. Moreover, SRS ensures the representativeness of an unbiased sample and enhances the accuracy of the inferences and generalisation about the population.

Besides, Stratified random sample offers several advantages over simple random sampling, include greater precision in estimation and support in-depth analysis of subgroup and variables. A stratified random sample can provide

greater precision than a simple random sample of the same size. Furthermore, our stratified sample can guard against an "unrepresentative" sample (e.g., excessive non-romantic samples from a mixed-romantic status population). From this, we can ensure that we obtain sufficient sample points to support a separate analysis of interested subgroups. Compared to simple random sampling, stratified random sampling, however, requires more administrative effort and the analysis is computationally more complex.

6 Conclusion and Limitations

While there are various factors that contribute to the educational gap in STEM among students of diverse social relations (e.g. gender, age, race and ethnicity), the present research suggests that romantic involvement can impact college students' math performance. This research thus provides an important step in identifying aspects of the interpersonal relationship in educational research and the situations of romance that interact to predict STEM outcomes. Limitations of this study include the absence of regression analysis to understand how romantic involvement positively or negatively influence academic achievement, and the lack of inclusivity of potential auxiliary variables (e.g. sex, parental education level). Consequently, our result may be generalised to our target population, but falls short with insufficient explanatory power and generalisation to a larger population of students of diverse social relations. Future research would benefit from investigating further the romantic elements, such as partner's preferences, partners' characteristics and relationships satisfaction, in shaping the academic achievement and aspiration of students of diverse social relations. From there, policy-makers and social programs can design and provide services in supporting youth to navigate interpersonal relationships, well-being and learning.

7 Appendix

7.1 R code for SRS, calculating estimates of mean, its standard errors and 95% confidence level

```
> grade <- read.csv("~/Desktop/student-mat.csv", header=
    TRUE)
> true.mean <- mean(grade$G3)
> true.mean
[1] 10.41519
> true.srs.prop <- length(which(grade$G3>=10))/N
> true.srs.prop
[1] 0.6708860
> true.prop.r <- length(which(grade$G3>=10 &
    grade$romantic=="yes"))/length(which(grade$romantic=="
    yes"))
```

```
> true.prop.r
[1] 0.6060606
> true.prop.nr <- length(which(grade$G3>=10 &
    grade$romantic=="no"))/length(which(grade$romantic=="
    no"))
> true.prop.nr
[1] 0.7034221
> N <- length(grade$G3)
> N
[1] 395
> n <- 195
> srs.indices <- sample.int(N,n,replace = F)
> srs.sample <- grade[srs.indices,]
> srs.ybar <- mean(srs.sample$G3)
> srs.ybar
[1] 10.482051
> srs.se <- sqrt((1-n/N)*var(srs.sample$G3)/n)
> srs.se
[1] 0.2448897
> srs.ci.mean <- c(srs.ybar-1.96*srs.se,srs.ybar+1.96*srs
    .se)
> srs.ci.mean
[1] 10.002067 10.962035
```

7.2 R code for SRS, calculating estimates of passing-exam proportion, its standard errors and 95% confidence level

```
> srs.prop <- length(which(srs.sample$G3>=10))/n
> srs.prop
[1] 0.671795
> srs.se.prop <- sqrt((1-n/N)*(srs.prop*(1-srs.prop))/n)
> srs.se.prop
[1] 0.023927
> srs.ci.prop <- c(srs.prop-1.96*srs.se.prop,srs.prop
    +1.96*srs.se.prop)
> srs.ci.prop
[1] 0.624897 0.718692
```

7.3 R code for Stratified Sampling Method, calculating estimates of mean, its standard errors and 95% confidence level

```
> romantic <- grade[grade$romantic=="yes",]$G3
> non.romantic <- grade[grade$romantic=="no",]$G3
> N.romantic <- length(which(grade$romantic=="yes"))
> N.nonromantic <- length(which(grade$romantic=="no"))
> n.romantic <- round(n/N*N.romantic,0)
> n.nonromantic <- round(n/N*N.nonromantic,0)
> n.romantic
[1] 65
> n.nonromantic
[1] 130
> romantic <- grade[grade$romantic=="yes",]$G3
> non.romantic <- grade[grade$romantic=="no",]$G3
> str.sample.r <- sample(romantic,n.romantic)
> str.sample.nr <- sample(non.romantic,n.nonromantic)
> str.FPC.r <- 1-n.romantic/N.romantic
> str.FPC.nr <- 1-n.nonromantic/N.nonromantic
> str.ybar <- (N.romantic/N)*mean(str.sample.r)+(N.
    nonromantic/N)*mean(str.sample.nr)
> str.ybar
[1] 10.557605
> str.var <- (N.romantic/N)^2*(var(str.sample.r)/n.
    romantic)*str.FPC.r+(N.nonromantic/N)^2*(var(str.
    sample.nr)/n.nonromantic)*str.FPC.nr
> str.var
[1] 0.0498121
> str.se.mean <- sqrt(str.var)
> str.se.mean
[1] 0.2231862
> str.ci.mean <- c(str.ybar-1.96*str.se.mean,str.ybar
    +1.96*str.se.mean)
> str.ci.mean
[1] 10.120159 10.99505
```

7.4 R code for Stratified Sampling Method, calculating estimates of proportion, its standard errors and 95% confidence level

```
> str.romantic.prop <- length(romantic[romantic>=10])/
    length(romantic)
> str.romantic.prop
[1] 0.6060606
> str.nonromantic.prop <- length(non.romantic[non.
    romantic>=10])/length(non.romantic)
> str.nonromantic.prop
[1] 0.7034221
```

```
> str.prop <- (N.romantic/N)*str.romantic.prop+(N.
    nonromantic/N)*str.nonromantic.prop
> str.prop
[1] 0.6708861
> str.var.prop <- (N.romantic/N)^2*(str.romantic.prop*(1-
    str.romantic.prop)/n.romantic)*str.FPC.r+(N.
    nonromantic/N)^2*(str.nonromantic.prop*(1-str.
    nonromantic.prop)/n.nonromantic)*str.FPC.nr
> str.var.prop
[1] 0.000568
> str.se.prop <- sqrt(str.var.prop)
> str.se.prop
[1] 0.023832
> str.ci.prop <- c(str.prop-1.96*str.se.prop, str.prop
    +1.96*str.se.prop)
> str.ci.prop
[1] 0.624175 0.717597
```

7.5 Reference

Brown, B. B., Bakken, J. P. (2011). Parenting and peer relationships: Reinvigorating research on family-peer linkages in adolescence. Journal of Research on Adolescence, 21(1), 153-165. doi:10.1111/j.1532-7795.2010.00720.x

Crosnoe, R. Elder, G. H. (2003). Adolescent friendships as academic resources: The intersection of friendship, race, and school disadvantage. Sociological Perspectives, 46(3), 331-352. doi:10.1525/sop.2003.46.3.331

Jeynes, W. H. (2005). Effects of parental involvement and family structure on the academic achievement of adolescents. Marriage Family Review, 37(3), 99-116. doi:10.1300/j002v37n03$_0$6

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. Developmental Psychology, 45(3), 850-867. doi:10.1037/a0014939

Quatman, T., Sampson, K., Robinson, C., & Watson, C. M. (2001). Academic, motivational, and emotional correlates of adolescent dating. Genetic, Social, and General Psychology Monographs, 127(2), 211-234.

Riegle-Crumb, C., & Grodsky, E. (2010). Racial-ethnic differences at the intersection of math course-taking and achievement. Sociology of Education, 83(3), 248-270. doi:10.1177/0038040710375689

Troncoso, P., Pampaka, M., & Olsen, W. (2016). Beyond traditional school value-added models: A multilevel analysis of complex school effects in chile. School Effectiveness and School Improvement, 27(3), 293-314. doi:10.1080/09243453.2015.1084010

UCI Machine Learning (2016). Social, gender and study data from secondary school students [Data file]. Retrieved from https://www.kaggle.com/uciml/student-alcohol-consumption.