Grade Bias Analysis

Shale Hunter

2021-12-02

Contents

Statistical analysis of grading biases

Introduction

The basic idea for this little project came to me when I was grading essays a few months ago for the class I am TAing this fall: as I methodically plodded through page after page of art history interpretation of *dubious* quality, I began to wonder if I were really treating each essay with equal care. By the time I graded my 50th and final essay, I was **sure** that I was not.

While I wasn't particularly happy with my intuitive conclusion that I probably wasn't being the fairest possible grader for my students, I was also a very busy masters student who wasn't going to spend an extra $15 min/essay * 50 essays = 12.5 hours$ to ensure that I was being 100% objective in my grading[1]. That said, the issue continued to linger in the back of my mind. So, when I was presented with the mandatory opportunity to conduct a self-driven statistical analysis for my Statistics for Environmental Data Science class, I jumped at the opportunity to tackle a question that had been nagging at me for several weeks, now secure in the knowledge that my curiosity would now be counted as progress towards my degree.

What follows is a statistical exploration of possible grading biases in 5 Teaching Assistants (including myself) and their 229 students who actually turned in essays.

Tidying the data

As TA, I have access to a whole lot of confidential student data which I probably shouldn't be sharing via a public blog post for any old internet traveler to discover and exploit. So I began by cleaning up the raw data both in order to protect student's privacy and to remove any unusable or missing data.

The cleaned dataset is available by request for anyone who wants to recreate the results of my actual statistical analysis with data that maintains the privacy of my students.

The basic components of the cleaned dataset are as follows:
- `name`: Student name (last, first initial), partially anonymized
- `pronouns`: Self-reported student pronouns (not present for all students)

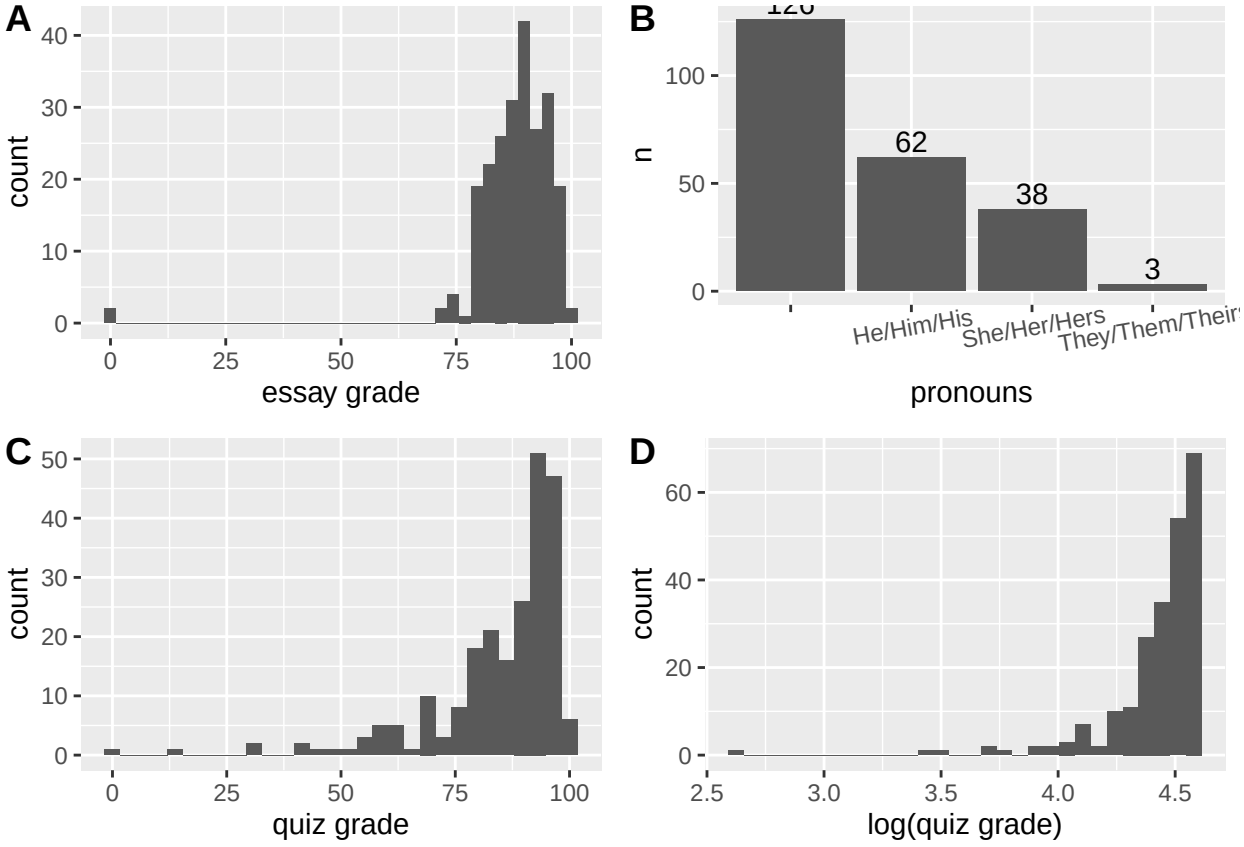
[1] After all, its an art history class, and an essay to boot - how much true objectivity is really expected here?

- `TA`: Teaching Assistant (Surname) and section, used for ordering students
- `essay`: essay grade, %
- `quizzes_pct`: cumulative quiz grade, %
- `participation_pct`: cumulative participation grade, % (not used in this analysis because some TAs had not finished entering participation grades at the time data was accessed)

name	pronouns	TA	essay	quizzes_pct	participation_pct
Barrientos C	He/Him/His	Hu]_10	85	60.19	0
Betova A	She/Her/Hers	Hu]_10	88	78.70	0
Bhaduri A		Hu]_10	88	86.11	0

Verifying basic assumptions of normality

Before getting started with answering my question, I'll take a quick look at the data to make sure it satisfies some basic requirements for normality. The basic histograms below show that essays (A) have a reasonably normal distribution (with one obvious low outlier). Quizzes grades (C), on the other hand, look like a log distribution might fit better - unfortunately, the log of quiz grades (D) also looks like it would fit a log normal curve (and so does the log of the log). So at this point I decided to just roll with the original quiz data with the recognition that any standard errors retrieved from this analysis will be effected, but bias should remain unaffected. Likewise, the pronouns data (B), which will be used in a follow-up analysis, have some clear limitations because of the uneven distribution of students who have listed their pronouns, particularly nonbinary students.



Autocorrelation

The goal of this analysis is to see if the team of Teaching Assistants (myself included) has been grading in a biased way. Specifically, because we/I graded the essay assignment in alphabetical order, it is possible to track if I graded a student's essay "in response" to the previous student's essay instead of based on the essay's individual merit. Hypothetically, this could take two forms: 1) differential grading, in which an essay is graded higher than it ought to be because the essay before was of particularly poor quality, or and essay is graded lower than it ought to be because the essay before was of particularly high quality. Or 2) uniform grading, in which an essay is graded higher/lower than it ought to be because it isn't all that different from the previous essay, and takes on a grade similar/identical to that of the previous essay. The null hypothesis here is that there is no effect by the previous essay's grade on the current essay, and therefore no grading bias (yay!).

To test this, we can use the `acf()` function, which compares each essay grade to each of the previous essay grades in `data_clean` up to `lag.max`. In this first example, we simply use a lag of 2 to compare each essay to the one immediately preceding it (calling a lag of 1 to the `acf()` function would only compare each value to itself, which of course will show perfect correlation).

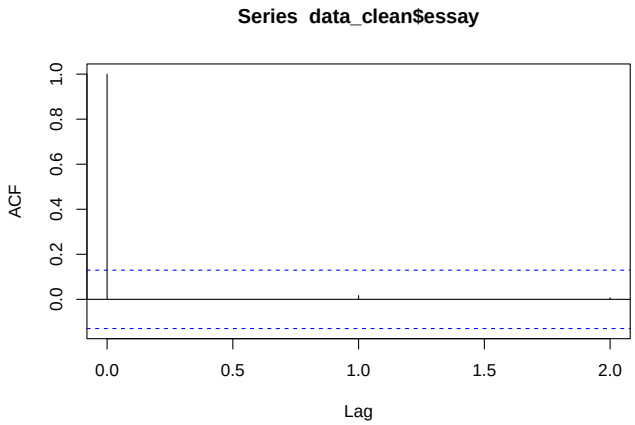


Figure 1: ACF plot with a lag of 1, no significant results.

As the autocorrelation plot above shows, there is no relationship between an essay's grade and its predecessor (as we would hope to see!). The dotted blue lines show the cutoff for statistical significance at a 95% confidence interval: because the autocorrelation value at lag = 1 is between 0 and the blue line (0.0167293 to be precise), it is statistically insignificant.

While my initial hypothesis was that essays graded immediately next to each other might be correlated, the autocorrelation performed above failed to reject the null hypothesis that there was zero correlation between grades. This is good news, but there are other ways we can use `acf()` to see if grading bias was introduced another way.

Next, we can use a larger lag in order to see if an essay's grade is correlated with grades of essays farther away on the roster:

Interestingly, this autocorrelation plot with a maximum lag of 10 shows that there are two points where an essay's grade is correlated with other essay's grades such that the acf value rises above the cutoff for statistical significance: t-4 and t-7.

While these results are unexpected, there are several possible explanations for why this might be the case: firstly, it is a possibility that all TAs graded essays in groups of 3-4 in order to avoid getting tired of grading and thereby giving lower grades to essays graded later on. If each grader came back from each break feeling more refreshed, happier, and more likely to give a better grade on an essay, then a pattern such as the one above might appear. However, it is important to note that at no point is there ever a significant negative autocorrelation, which seems to argue against a pattern like this - if an essay's placement at the beginning of

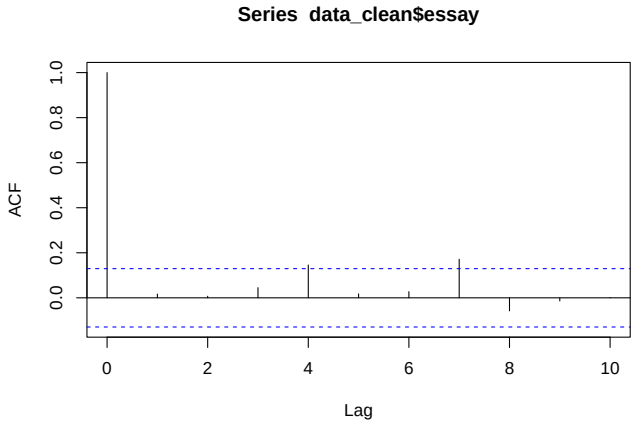


Figure 2: ACF plot with lag of 10, significant results at lag = 4, 7 (indicated by passing the dotted blue line).

such a group lead to higher grades, then another essay's placement at the end of the group should likewise indicate lower grades, and a negative acf value. But, that is clearly not the case in this data.

An alternative explanation is that TAs may be consciously or unconsciously tracking the numbers of A's, B's, or C's they have given, and compensating for higher or lower grades every few essays. For example, if a TA hadn't given any A's for several essays in a row, they might give the next essay an A in order to make up for a perceived deficit (even if the essay in question only really deserved a B+).

It should be noted that I ran the `acf()` function with a lag as high as 50 and no other significant correlations were found, so even if either of these biases were real the periodicity does not extend beyond a lag of 7.

Gender Bias

Beyond my initial curiosity regarding the possibility of an autocorrelation bias in my grading, I was also recently reminded of the potential for gender bias in grading. So this next section will take a look at the subsection of students who have reported gender information on the class portal in order to identify a possible systemic devaluation of academic work based on gender.

For this test, we first need to restrict our observations to only those students who have provided pronouns.

Then, we can conduct a simple linear regression with the `lm()` function to see if there is an effect of gender on essay grades:

Interestingly, the model gives a positive coefficient for women (1.2580465) but a negative coefficient for nonbinary (-2.5374933) in relationship to a male default. This means that the model expects women's essay grades to be higher than men's, but nonbinary students' essays grades to be lower.

There are two major caveats to these results: firstly, neither of the gender coefficients had p-values even close to the traditional cutoff of 0.05 (she/her: 0.2728258, they/them: 0.4332378). Furthermore, with only 3 students listing their pronouns as They/Them/Theirs, it is particularly difficult to make stastically sound assumptions about this set of students even in relation to the rest of this dataset.

Although none of the gender coefficients are significant (indicating a lack of any conclusive grading bias based on gender), the `quizzes_pct` coefficient of 0.1111361, which indicates that for every one percent increase in a student's overall quiz grade their essay grade is expected to go up by 0.1111361 percentage points, is statistically significant at the 0.05 level given a p-value of 0.0118374. This validates my choice of quiz grade as a good benchmark against which we can test for bias, because it shows that a student's own grades are correlated with each other (as they should be).

All of this can be visualized with the plot below:

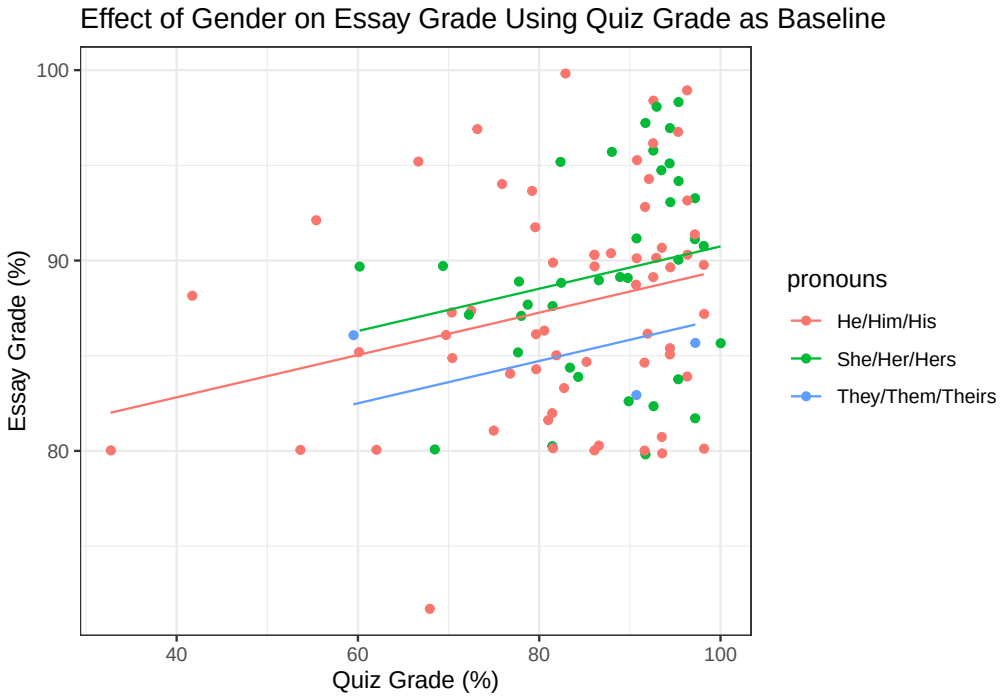


Figure 3: Essay grades are marginally higher for women and lower for nonbinary students as compared to men, but not statistically significant.

Visualizing the model with the plot above makes it clear that there is a positive relationship between quiz grade and essay grade, as indicated by the positive and statistically significant `quizzes_pct` coefficient in the model. Likewise, though the offset of the lines for different genders shows women with overall higher grades and nonbinary students with overall lower grades than men, the large amount of spread in all the data points is a visual indicator that these results are not statistically significant.

Conclusion

After conducting this analysis, I can more or less rest easy knowing that I am as honest a grader as can be expected. A linear model showed that essay grades were correlated with the same student's quiz grades, but not significantly correlated with gender. Similarly, a test of autocorrelation showed that the grade of the essay immediately beforehand has no effect on an essay's grade, which was my original concern. And while two essay lags were found to have significant correlation, this can be attributed to the type of human failings which are all too common, and are only being properly understood as we gather and interpret increasingly large amounts of data on the topic.