RateMyProfessor Data Analysis

Preprocessing
Since it is advised to handle the issues of a confounding factor - teaching experience - and low sample size, we decide to address them by selecting certain data and dividing them into groups:

To address the issue of low sample size, we believe that there is not enough information for us to conclude anything meaningful about a professor with less than 5 ratings, therefore for this part of the project we drop the data of these professors. To control the confounder, we select professors who received 5 to 24 Number of Ratings and group them into intervals of 5 ratings each (5–9, 10–14, 15–19, 20–24). We assume that professors with similar numbers of ratings will have similar levels of experience. We perform significant testing individually for each group.

Also, since we are tasked with finding the difference in reported ratings between male and female professors, we are only interested in professors who are either male OR female as shown by the data. We drop data for professors with (Male=1, Female=1) and (Male=0, Female=0). And, to avoid collinearity, we drop the column “Male” and classify the gender of the professors by their record in column Female.

Question 1

Activists have asserted that there is a strong gender bias in student evaluations of professors, with male professors enjoying a boost in rating from this bias. While this has been celebrated by ideologues, skeptics have pointed out that this research is of technically poor quality, either due to a low sample size –as small as n = 1 (Mitchell & Martin, 2018), failure to control for confounders such as teaching experience (Centra & Gaubatz, 2000) or obvious p-hacking (MacNell et al., 2015). We would like you to answer the question whether there is evidence of a pro-male gender bias in this dataset. Hint: A significance test is probably required.

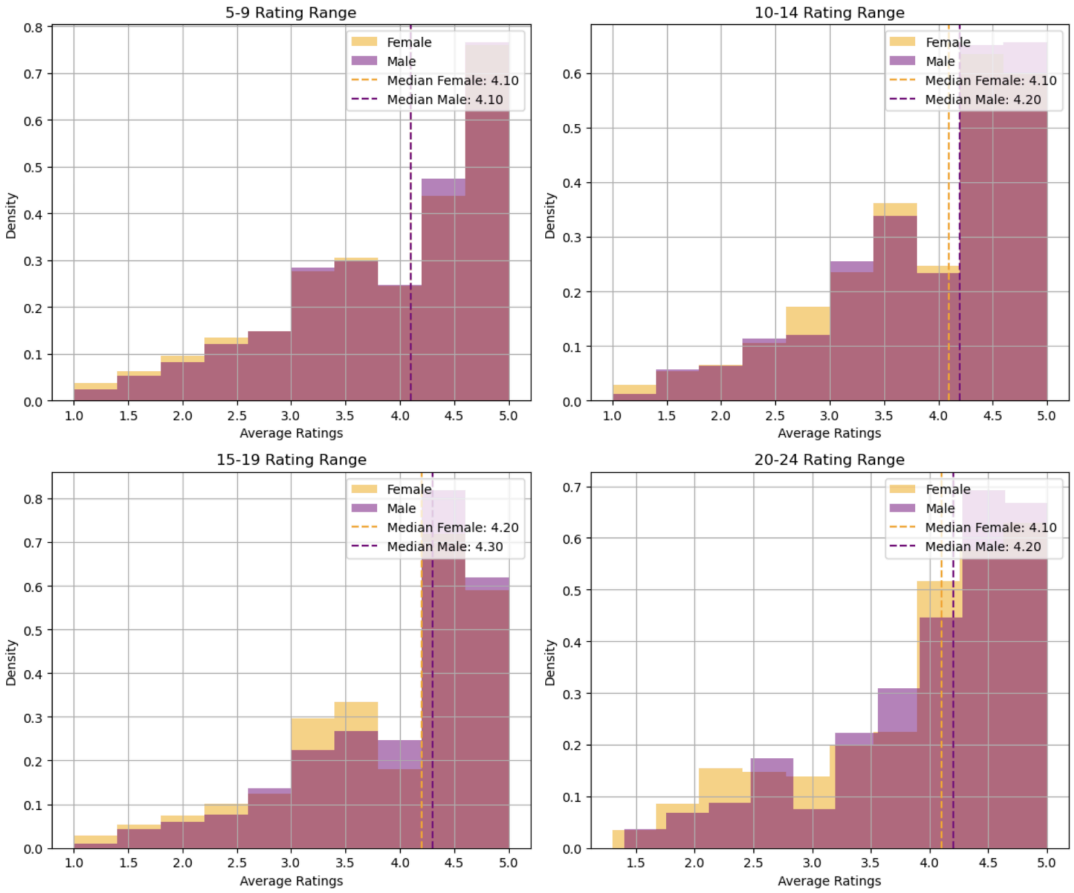
Assumptions: see above
How they were addressed: see above

Model Justification:
Since we are comparing the ratings, it might not be wise to calculate the mean: the unit difference is not constant. Therefore, we consider it more appropriate to compare by the median. Hence, Mann-Whitney U test is our model of choice.

Null Hypothesis:
Our null hypothesis is that there is no significant difference between the average rating received by female and male professors.

Test result:

We run Mann-Whitney U test for each group with a certain number of ratings received (5 to 9, 10 to 14, 15 to 19, and 20 to 24) to compare the gender difference between Average Ratings while controlling for experience. All 4 results are insignificant, indicating that while controlling for the confounder - experience, in this case - there is no significant gender bias for students' evaluation of the professors.



In the visualizations, it's indicated that for each group the gender difference in Average Ratings between male and female professors are not very prominent, and our test results validate this statement.

```
All Results:
  Rating Range  U Statistic   P Value
3          5-9   15623499.0  0.066782
0        10-14    1678511.0  0.059399
1        15-19     276132.5  0.064070
2        20-24      66648.5  0.198804
```

Question 2

Is there a gender difference in the spread (variance/dispersion) of the ratings distribution?

Again, it is advisable to consider the statistical significance of any observed gender differences in this spread.

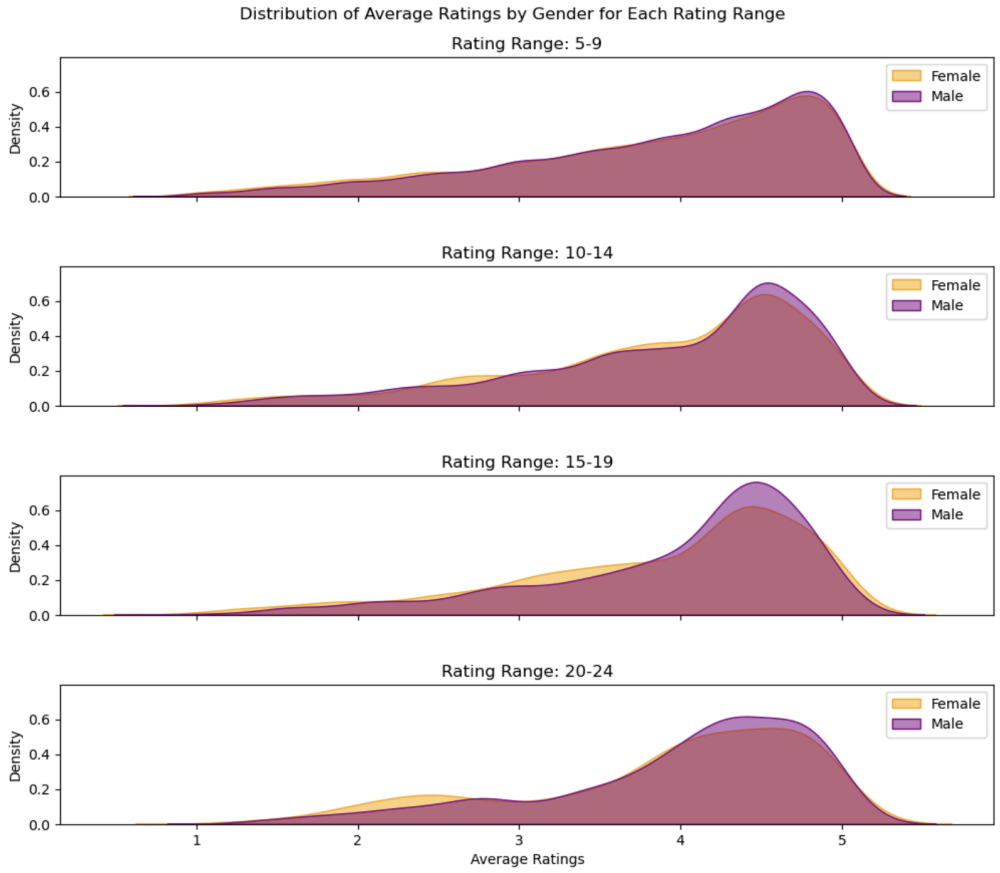
Assumptions: see above
How they were addressed: see above

Model Justification:
Since we want to test if there is a difference between the spread/distribution of the average ratings received by female and male professors, the Kolmogorov-Smirnov test would be a suitable one to use.

Null Hypothesis:
Our null hypothesis is that there is no significant difference between the distribution of the average rating received by female and male professors.

Test result:
We perform KS tests for each group to compare the gender difference between the distribution of Average Ratings while controlling for the experience as potential confounders. All four results are insignificant, indicating that while controlling for the confounder, there is no significant

gender bias for the distribution of students' evaluations (Average Ratings) of the professors.



The plot does not indicate a significant difference between the distributions of Average Ratings received by female and male professors, aligning with the results of KS tests for all groups.

```
All Results:
  Rating Range  KS Statistic   P Value
3          5-9      0.021408  0.148299
0        10-14      0.036947  0.153615
1        15-19      0.078339  0.017339
2        20-24      0.056481  0.571946
```

Question 3

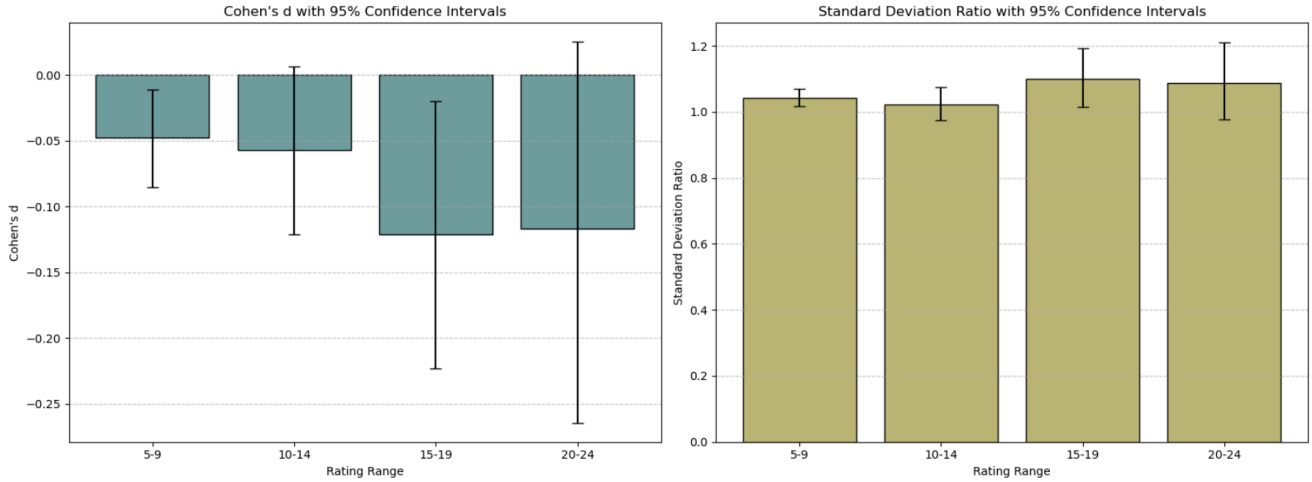
What is the likely size of both of these effects (gender bias in average rating, gender bias in spread of average rating), as estimated from this dataset? Please use 95% confidence and make sure to report each/both.

Assumptions: see above
How they were addressed: see above

Effect size estimation:

It's not ideal, but since we have not covered how to estimate the effect size between the difference of the median, we resolve to use Cohen's d for estimating the effect size for gender difference between Average Ratings professors received. It will still return accountable results, since in the original research the researchers have already taken the mean of the ratings and believed them to be meaningful (Average Ratings). For estimating the difference between how the Average Rating is distributed, we use the ratio between the standard deviations of female and male professors' Average Ratings. If the distributions are similar for two groups, the ratio should be around 1.

Result:



Bootstrap Results:

Rating Range	Cohen's d	Cohen's d CI Lower	Cohen's d CI Upper	Std Ratio	Std Ratio CI Lower	Std Ratio CI Upper
5–9	−0.047640	−0.085092	−0.011212	1.042182	1.016656	1.069233
10–14	−0.056990	−0.121462	0.006633	1.022628	0.973538	1.075076
15–19	−0.121099	−0.222892	−0.019789	1.098763	1.015683	1.191891
20–24	−0.116916	−0.264550	0.025152	1.086718	0.977847	1.208940

For each group, we estimate the effect size by calculating the empirical Cohen's d and the ratio between two standard deviations. Then, we use the bootstrap method to calculate the 95% confidence interval of the effect. As we can see in the result and more clearly from the visualization, the difference in Average Ratings and their distributions is larger for groups with more number of ratings received, and the variance is also larger, reflecting in wider confidence intervals. However, none of the effects is really prominent, for their numerical values are all rather negligible.