Mier Chen

Mc5wx

# Dr. Nuhfer’s Data Analysis

## Executive Summary

## Introduction

Dr. Nuhfer brings us a data set generated from a survey he designed.

**Survey Deisgn**: It includes a measure of demonstrated competence of science literacy from a test instrument and four measures of self-assessed competence, one as a composite score from a 25-item knowledge survey instrument and three from single questions. I rename the self-assessed competence as global\_rating1, global\_rating2, global\_rating3 and actual\_KSSLCI.

Our client Dr. Nuhfer is interesting in the following questions:

1. Any other useful variables I should include in the survey?
2. Is the size of data large enough and representative?
3. How to distinguish the participants who engaged in random guessing?
4. Does our data exhibit support for a Dunning-Kruger Effect in our study populace?
5. Do our data support the contention that most people are overly optimistic?
6. Is there any evidence in our data that supports that experts are actually any better than novices in self-assessment?

In this data set, we have a total of 1154 observations and 82 variables. However, if we want to group some variables and evaluate each group, we may need more data to support that. For example, we don’t have enough observations who are professors and graduate students.

Also, I find that 504 observations are from Institution A and 414 students are from Institution D. The rest of institutions has only less than 100 observations. I suspect that it may cause some bias in our model.

## Approach

1. **Cleaning the data and recoding the variables**Since I will use all the demographic variables, from gender to institution, I drop the observations with any missing value in these columns. The total number of observations after cleaning is 1125. I also recode the categorical variables to numeric numbers. It not only accelerates the computation, but also makes later machine learning normalization easier.
2. **Explore some important variables**, such as gender, class rank, science major and institution. Check if that variable is significant.
3. **Machine Learning Method: SVM**

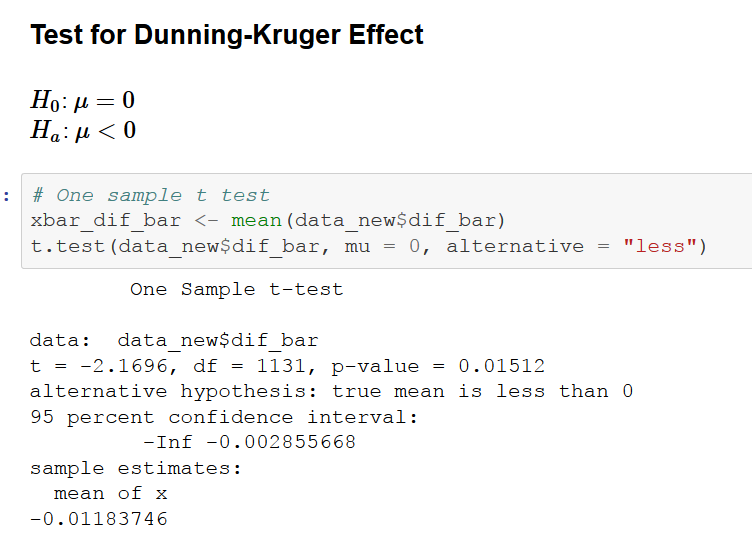
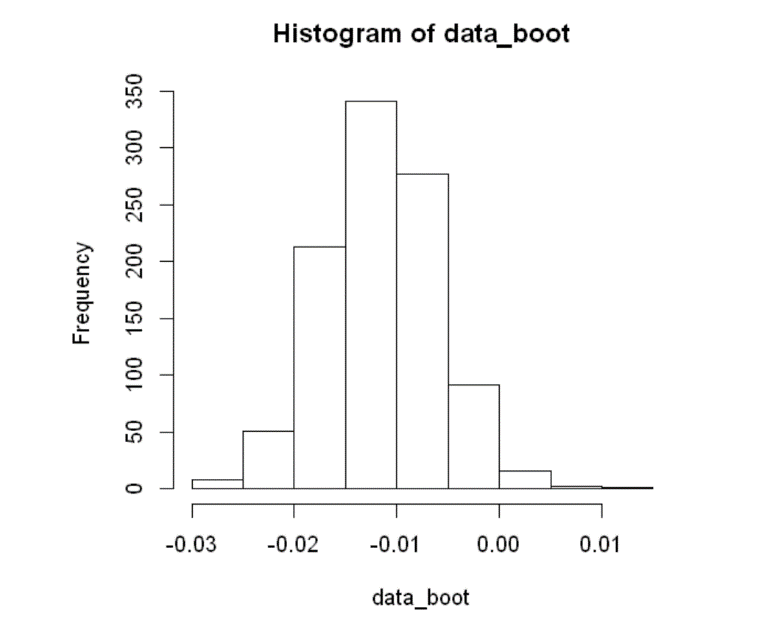
## Result

#### Calculating the

#### 

I create two matrixes, one for the knowledge test records(K) and the other for the actual test records (S). I use the above method to create . Then, we calculate the mean of all and obtain that individual’s dif\_bar. If the dif\_bar is smaller than 0, we can say that he/she **overestimate** themselves. If that person’s dif\_bar is higher than 0, we can say that he/she overestimate themselves. **This method is different from Dr. Nuhfer’s way of calculating d. It decreases many noises in the process.**

Since the qq plot for dif\_bar doesn’t look good. We use bootstrap to sample 1000 points from the data and draw another histogram. From the histogram, we can see that the mean is around -0.01. We can conclude together with the t test that people are more likely to overestimate themselves. It answers the fifth question that our client asked.



For the following reports, I only select some data visualizations that I think it the most representative. For the full report with code, please refer to the electronic version for details.

#### 2. Gender & Class Rank

## 

From the boxplot, we can easily see that the female students are more likely to underestimate themselves. We found more male students who overestimate their score a lot, while several more female students much underestimate their scores. It is interesting to see that we have female senior students who much underestimate their scores as well as much overestimate their scores.

We don’t have enough data to support that the professors and graduate students have significant difference in gender due to the lack of data.

#### Class Rank for undergraduate students only

## 

We have enough data records for all four years of undergraduate students. Hence, the distribution of dif\_bar for the undergraduate students is more persuasive than graduate students and professors. From the density plot above, we can see that the fourth-year students have a more distinguish distribution than the other three levels’. We also observe a small peak on the bottom-left of the graph. It indicates that the senior students are more likely to get a correct estimation than the other three levels.

#### Heatmap for Gender & Class Rank

## 

The orange color represents that people in that group overestimate themselves, and the blue color represents that people in that group underestimate themselves. The saturation of colors represents the level of the over or under estimation.

From the heatmap, we can easily find that the second and third year male students are more likely to overestimate themselves. I hesitate to conclude that the female professors and female graduate students are more likely to underestimate themselves since we don't have lots of data to support it. Moreover, since their scores are pretty high, they have less room to overestimate themselves.

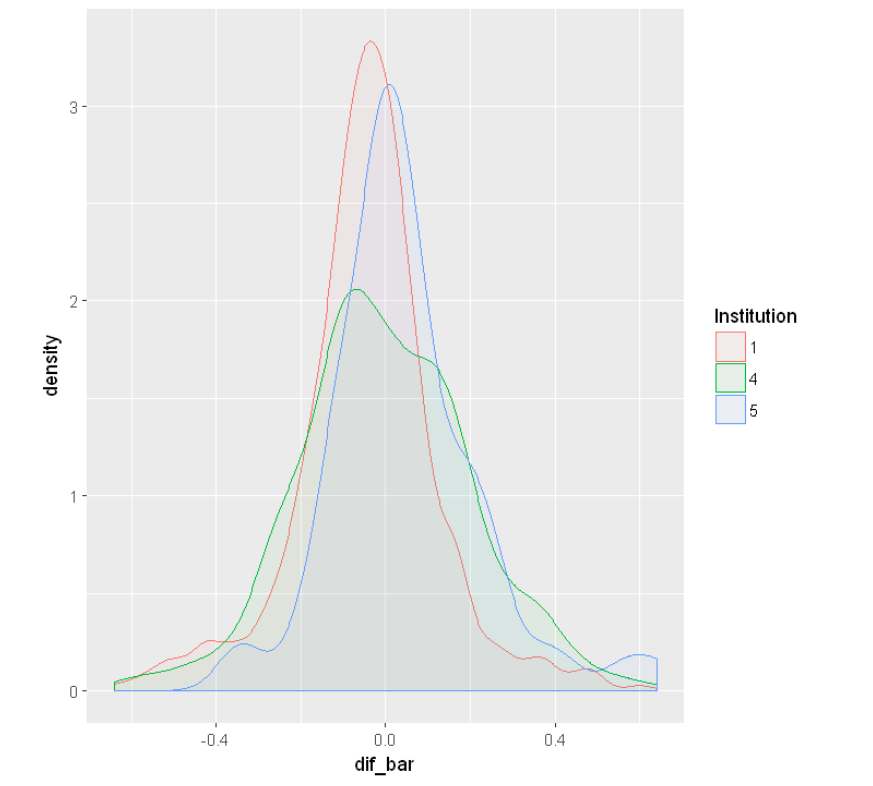
#### Heatmap for Gender & Institution

## 

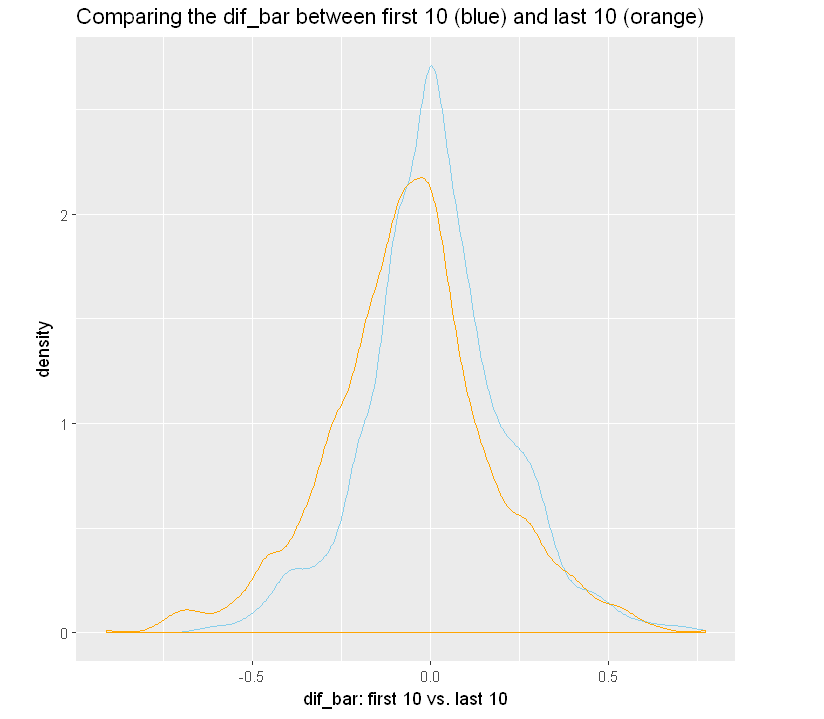
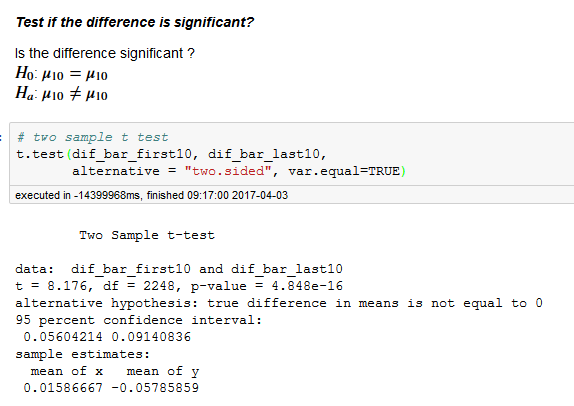
Also, our client mentioned that the institution may play a role in the mean of differences. Hence, I create another heatmap to observe the two gender groups in each institution.

It is interesting to find that we don't have any group that typically overestimate themselves. The male students in the Institution 4 and Institution 1 slightly overestimate their ability. It is also interesting to see that Institution A is the only school that has both male and female students overestimate themselves.

#### Density plot for Institution A, D and E

Since we only have over 63 observations for these three institutions, I think it makes sense to only plot them three. The density plot above indicates that Institution E has a very different distribution comparing to Institution A and D. It may worth digging the reason behind it.

#### First 10 questions vs. Last 10 questions

When I was doing the survey, I gradually lost my patience after finishing the first ten questions. I think it may be interesting to see if there is a significant difference between the first 10 and the last 10 questions. I drop the middle 5 question because the 15th question is a pretty tough one. It may cast some bias.

From the graph and the t test, we can conclude that the difference between the first 10 questions and the last 10 questions is significant at 95% confidence level. It may give us some inspiration in detecting random guessing.

#### Machine Learning Results:

**Steps:**

1. normalize the variables

2. create the personality variable: underestimate, overestimate, normal

3. random sample a training data set, 600

4. check each model's training error

5. evaluate the model with least training error, check the accuracy rate

I tried polynomial model, Gaussian model, linear kernel, Laplace dot model and Laplacian kernel and check their training error. Among them. The Gaussian model has the lowest training error of 0.45. After evaluating the model with testing data, the accuracy rate is around 0.51, which is higher than random guessing, 0.33. However, this method works, but doesn’t seem to be very effective.

## Conclusion

After using bootstrap, we can conclude that people are more likely to overestimate themselves. Our data exhibit support for a Dunny-Kruger Effect. It would be better if we can collect more data of professors and graduate students to be more confident.

Gender, class rank and science major are significant factors for dif\_bar. Also, the difference between the first 10 and last 10 questions is significant.

We may need more data for professors and graduate students in order to make better conclusion.