Academic Achievment of Asian and non-Asian Graduates of Williams College

by Jackson Moss '19

Abstract A recent NY Times article reported that College-bound seniors taking the SAT divide harshly along lines of race, especially those on the high end of the distribution. The **wgd** package uses publically available data from the Williams College Registrar to load and determine statistics about the names and academic achievements of graduating classes between 2004 and 2016. The package uses surname analysis to estimate whether a student is Asian or not Asian, and then computes statistics and graphics comparing the two groups.

Introduction

A student's SAT score determines in large part which colleges he/she will get into and which colleges he/she will not. At a highly selective school like Williams, most competitive students boast SAT scores in the top 5% of all test-takers. Top SAT scores, those between 750 and 800, are highly skewed by race. In fact, 60% of these top scores can be attributed to Asian Students. In the class of 2015, the college reported that 13.5% of American students were of Asian descent. Assuming that half of the 7% of international students are of Asian descent, we can estimate that only 17% of all Williams graduates since 2004 have been Asian.

Asian students are 60% of those scoring above 750 on the SAT, but represent only 17% of the Williams population. This might suggest that, on the basis of SAT scores, Williams discriminates against Asian applicants. Namely, Williams college admissions might have a higher SAT score standard for Asian applicants. While SAT scores are not a perfect predictor for academic success in college, we might expect to see this discrepancy in graduation honors data. The **wgd** package uses surname analysis to estimate whether a student is Asian or not Asian, and then computes statistics comparing the two groups with respect to academic achievement data.

Data

The raw data used for this analysis is in the form of PDFs posted to the website of the Office of the Registrar of Williams College. The data was converted into text files using PDFelement, an online PDF to text editor. After the data is read into R, we separate into columns in order to construct a meaningful dataframe. We first split based on the first space, separating the first name of the student from the rest of the information. We then separate by the instance of the comma, which splits based on whether or not the student received subject honors.

```
y <- colsplit(df," ",c("firstname","rest")) # separate first name
x <- colsplit(y[, 2], ",", c("lastname", "rest")) # separate last name block
split.x <- colsplit(x[, 1], "with", c("name", "honors")) # split up based on honors
full <- as.data.frame(cbind(y[, 1], split.x[, 1], test), stringsAsFactors = FALSE)
```

We now have a dataframe summarizing the first and last names, and information about subject honors. We then remove instances where the last name has an ending that would corrupt the data, like "Jr." or "III". Next, we split the last column of the data based on the type of subject honors the student received.

```
split.complete3 <- colsplit(complete3[, 3], "honors", c("highest", "regular"))
```

In each dataset, we have four lines representing the four levels of Latin honors (Summa Cum Laude, Magna Cum Laude, Cum Laude, none). Each of these lines begins with the word "Bachelor" and follows the order from highest Latin Honor to no Latin Honor. We thus create a new column based on the locations of these four lines.

```
find <- grep("Bachelor", xy[, 1])
  summa <- rep("summa", find[2] - 1)
  magna <- rep("magna", find[3] - find[2] - 1)
  cum <- rep("cum", find[4] - find[3] - 1)
  none <- rep("", nrow(xy) - find[4] - 1)
  bind <- c(summa, magna, cum, none)
```

After cleaning up extra spaces and special characters, we perform a similar analysis to create a column reporting the year that the student graduated. We also add columns containing the number of characters in each students first and last name. After combining the relevant columns, we have an eight column dataframe that retains most of the information that the college provides.

```
#>   year latin honors first name length  last name length subject honors
#> 2 2016        summa   Benjamin      8 Augenbraun     10        highest
#> 3 2016        summa      Julia      5     Damion      6
#> 4 2016        summa      Jesse      5    Freeman      7        highest
#> 5 2016        summa       Emma      4 Harrington     10        highest
#> 6 2016        summa      Lydia      5  Heinrichs      9        highest
#> 7 2016        summa     Isaiah      6    Leonard      7        highest
#>       subject
#> 2     Physics
#> 3
#> 4 Mathematics
#> 5     Science
#> 6     English
#> 7     Science
```

Surname Analysis

Surname analysis in order to judge whether or not a student is Asian is an important part of this analysis. Inside the **nameSummary()** function is a surname analysis function that guesses whether a student is of Asian descent or not. The function tests whether the student's surname matches with any of 378 common Asian surnames. The list of names was taken from a 2010 Canadian surname validation study. While the complete study involved over 9,900 names, only the top 378 were available. The study reported a positive predictive value for Asians at just over 90%. Since only the most popular names are used in this analysis, we expect a much lower positive prediction rate. This analysis reports that 245 of the 6875 students who have graduated since 2004 are Asian. Assuming that the true value is 17%, the positive prediction rate is 21.0%.

It is important to discuss how this low positive prediction rate could affect the results. The full data set is split between students who are guessed to be of Asian descent (Asian data set) and students that are guessed to not be of Asian descent (Other data set). Our Asian data set contains 245 students who are highly likely to be Asian. There is no reason to believe that Asians with more common surnames perform differently at Williams than Asians with less common surnames. While a higher positive prediction rate and larger sample sizes might give us more reliable estimates, our estimates for proportions for our Asian population should be accurate. Our other data set contains 6630 students of all races, some of which are Asia. In fact, we guess that 79% of Asians who have graduated from Williams are still contained in this data set. This is attributed with 13.4% of the other data set. Assuming that Asian students have a higher likelihood of earning Latin and Subject Honors, we thus expect our other data set to overestimate the true proportion of non-Asian students receiving academic honors. Consequently, we expect our estimates of the differences between the two groups to be an small underestimate.

Use EphData()

The information in the locally stored text files can be loaded and parsed using the **EphData()** function. The function creates an 8 column dataframe containing information about Williams College graduates for year that the text file refers to. The function has one argument and is used as follows.

```
EphData("dataset")
```

Use nameSummary()

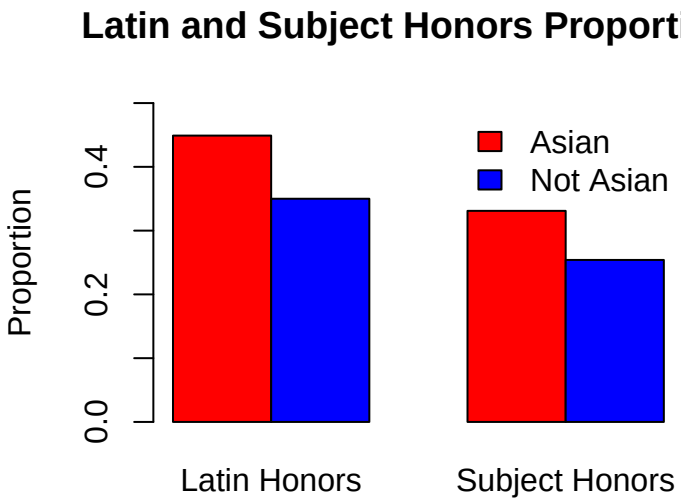
The function **nameSummary()** is used to generate statistics and graphics about Williams College graduates since 2004. The function separates the full data set between students guessed to be of Asian descent and students guessed to not be of Asian descent. It then performs analysis on the differences between these groups with respect to the likelihood of earning academic honors and the length of first and last names. **nameSummary()** has one argument and can be used as follows.

```
nameSummary("type")
```

Data Analysis

As stated earlier, there could be reason to believe that Williams admissions discriminates by race in part on the basis of SAT scores. Using the **nameSummary()** function, we can split up the data into an Asian student group and a non-Asian student group. We can then calculate the proportions for students from each group earning academic honors.

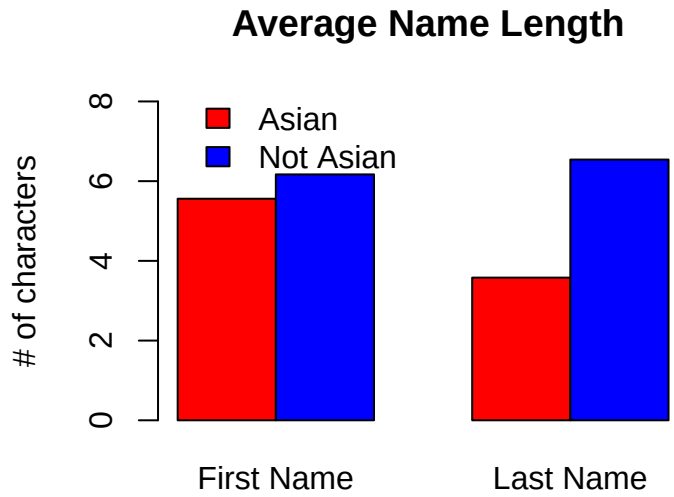
```
nameSummary("honors")
```



We see that Asian students, as estimated by the surname analysis, are more likely to earn both Latin Honors and Subject Honors. In fact, Asian students are 9.9% more likely to earn Latin Honors and 7.7% more likely to earn Subject Honors. As distressed earlier, we expect that this difference is a slight underestimate of the true difference between Asian students and non-Asian students at Williams college.

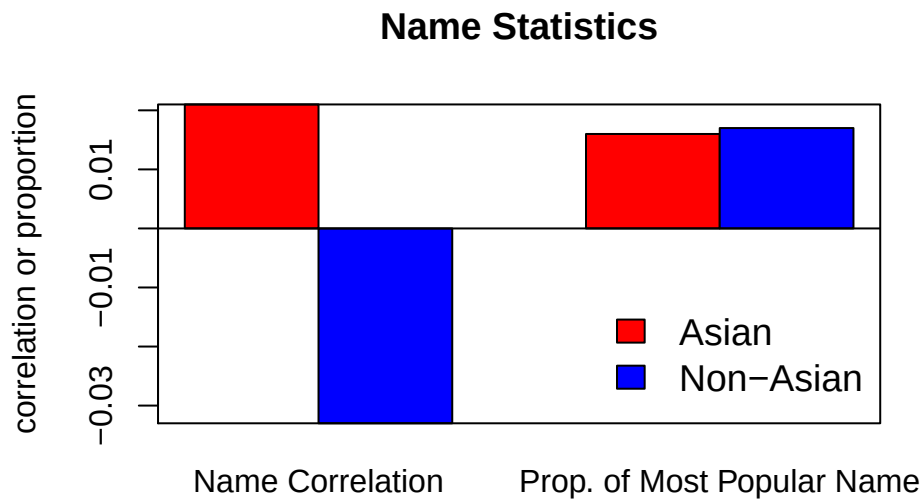
Another interesting question is whether the students estimated to be Asian have longer or shorter names than the students who are estimated to not be Asian. While it is well known that Asians tend to have shorter surnames, do they also tend to have shorter first names?

```
nameSummary("length")
```



We see that Asian students tend to have shorter first names and shorter last names than non-Asian students. First names tend to be 0.61 characters less for Asians than non-Asians, and surnames tend to be 2.96 characters less for Asians than non-Asians. Similarly, these differences are expected to be slight underestimates of the true differences between the populations. However, part of this difference could be do to the surname analysis. Since only the top 378 names were used, we are only including Asians at Williams with common last names. Less common names usually tend to be longer, so we are in effect probably not including the Asians with longer names. It is unclear how much this bias is affecting the results.

```
nameSummary("popular")
```



We see that there is a very small positive correlation between the length of first names and the length of last names for Asians at Williams. This drops to a small negative correlation for non-Asian students. The second metric reports the proportion of total students that the most popular name represents. These metrics are almost identical, at 1.6% for Asian names and 1.7% for non-Asian names. This would suggest that the distribution of first names of Asian students and first names of non-Asian students are equally diverse at Williams.

Conclusion

Our hypothesis can be restated as follows. Asian students are 60% of those scoring above 750 on the SAT but only 17% of the Williams population. This suggests that Williams admissions might have a higher standard for Asian applicants with respect to SAT scores. If so, we might expect to see this discrepancy in graduation honors data. This analysis concludes that there are marked differences in academic achievement between students this analysis estimates as Asian and students that this analysis estimates as not Asian. In regards to the question posed in the introduction, we see that Asian students, as estimated by surname analysis, are more likely to earn both Latin Honors and Subject Honors. In fact, Asian students are 9.9% more likely to earn Latin Honors and 7.7% more likely to earn Subject Honors.

While we have showed that there is a difference between the two populations as estimated by surname analysis, more research needs to be conducted in order to reliably conclude that Williams admissions discriminates against Asians in part on the basis of SAT scores. Williams takes a holistic view of student applications, considering many diverse factors such as test scores, grades, extracurricular activities, service, and geographic location. Unfortunately, Williams College is not likely to provide a data set containing this information.

However, the fact remains that Asian students are more likely to earn Latin Honors than non-Asian students. We also estimate that the differences reported in this study are small underestimates of the true differences between students of the two groups. Better surname analysis would decrease the number of Asians in our non-Asian data set and thus decrease bias. For example, the Canadian study that used over 9,900 names reported a positive predictive rate of over 90%. This analysis estimates a positive predictive rate of 21.0%. Further, coupling this analysis with census data could indicate not only whether a student is Asian or not Asian, but suggest what country or region the student is from. Breaking up the data into more groups could provide more reliable estimates of the differences between students.

Regardless, the **wgd** package includes functions that allow for easy reading and analysis of Williams College graduate data. Depending on the format, this analysis can be generalized to other institutions that publish data in a similar format. Further, and more importantly, the dataframes created using the EphData() function can be used to conduct analysis other than the one discussed here.

Jackson Moss '19
Williams College
39 Chapin Hall Drive
2964 Paresky
jm21@williams.edu