Comparision between Female and Male Proportions of Williams Students Who Graduated with Latin Honors

by Panchanok Jumrustanasan '19

Abstract The package compares the proportions of female and male students who graduated with Latin honors from class of 2003 to 2016 using **gender** package.

Introduction

The gender distribution of first-year students are approximately equal between male and female students(49 percent male and 51 percent female students for the class of 2019.) Even though the slightl difference in numbers indicates some sort of irrelavance in gender and acceptance, the impact of Williams College settings on adacemic standing of students of each gender is worth studying.

The paper takes Latin honors students received as an indicator of academic achievement of students. It focuses on the proportions of students of each gender from each class year from 2003 to 2016 who graduated with Latin honors. The gender is determined using **gender** package(Mullen, 2015).

Proportions

In this paper, *proportion* refers to the proportion of Williams students of each gender who graduated with Latin honors. The number is obtained by dividing the number of students who graduated with Latin honors by the total number of students of that gender in the same class year. The proportions of female and male students are compared in this fashion to prevent the interferance from the fact that students of one gender are more likely to receive Latin honors than the other if there are more students of that gender.

Raw Data

The information of Williams College graduating classes were from Williams College Bulletin pdf files available on the Office of the Registrar of Willaims College website (Wil, 2016). The files were converted into text files using online file converter and manually modified. At this point, the files contain only the information of graduating students, but still need more manipulation from the functions in the package.

getHonorName()

`getHonorName()` extracts information from the section indicated by Latin honors. It takes 2 arguments: 'filename' and 'honor.' 'filename' can be the file name of the graduating class information in the package or a dataset in the package. File names start from 't0203.txt' for the class of 2003 to 't1516.txt' for the class of 2016. 'honor' tells the section that information should be taken from. They can be '`all`', '`summa`', '`magna`', '`cum`', and '`none`'.

The returned value is a dataframe with one column of Latin honor and one column of other information that will be cleaned later.

```
honor.data <- getHonorName("t0203.txt", "all")
honor.data[1:15,]
```

```
#>                                          dat_honor       honor
#> 1                 Bachelor of Arts, Summa Cum Laude Latin Honor
#> 2  †*Emily Patricia Balskus, with highest honors in Latin Honor
#> 3                                          Chemistry Latin Honor
#> 4                              *Aimee Rose Candelore Latin Honor
#> 5                            *Megan Elissa Delehanty Latin Honor
#> 6       *Katherine Keleher Desormeau, with highest Latin Honor
```

```
#> 7                        honors in Literary Studies Latin Honor
#> 8                             *Kristina Gray Fisher Latin Honor
#> 9                           *Christopher Edward Goggin Latin Honor
#> 10  *Johanna Dorothy Heinrichs, with highest honors Latin Honor
#> 11                                          in Art Latin Honor
#> 12                             *Bradley Thomas Howells Latin Honor
#> 13      †Theresa Cunningham O'Brien, with honors in Latin Honor
#> 14                                       Biology Latin Honor
#> 15                                *Julia Ann Snyder Latin Honor
```

cleanData()

The dataframe from `getHonorName()` might contain some elements such as section headers, non-names, and some special marks. Working with internal helper functions, `cleanData()` detects and gets rid off these elements, leaving only vital elements that represent students.

It takes a dataframe from `getHonorName()` as its only argument. The function separates first names from middle and last names. The gender column is also added to the dataframe.

```
clean.data <- cleanData(honor.data[1:15,])
clean.data
```

```
#>      firstname       mid/lastname       honor gender
#> 1        Emily   Patricia Balskus Latin Honor female
#> 2        Aimee     Rose Candelore Latin Honor female
#> 3        Megan   Elissa Delehanty Latin Honor female
#> 4    Katherine  Keleher Desormeau Latin Honor female
#> 5     Kristina        Gray Fisher Latin Honor female
#> 6  Christopher      Edward Goggin Latin Honor   male
#> 7      Johanna  Dorothy Heinrichs Latin Honor female
#> 8      Bradley     Thomas Howells Latin Honor   male
#> 9      Theresa Cunningham O'Brien Latin Honor female
#> 10       Julia         Ann Snyder Latin Honor female
```

However, due to the limitation of **gender** package, the gender of some names, especially non-English names, are NA.

ratio()

`ratio()` takes a dataset from '`wstudent.xxx`' series (see more details in '`Datasets`' section below) in the package as its only argument. It returns a table of the proportions of that input.

```
ratio(wstudent.three)
```

```
#>
#>    female      male
#> 0.3903509 0.3122530
```

Datasets

The package provides ready-to-use datasets; they are datasets in '`wstudent.xxxx`' series and '`all.ratio`'.

'`wstudent.xxxx`'s are the clean manipulated version of text files that are saved as datasets within the package. A dataset in this series contains all students in the class year, their genders, and Latin honors they received. To use the datasets, call `data()` on their names from the list: '`wstudent.three`', '`wstudent.four`', '`wstudent.five`', ..., '`wstudent.sixteen`'. For example, to get the dataset for class of 2003,

```
data(wstudent.three)
wstudent.three[1:15,]
```

```
#>      firstname       mid/lastname           honor gender
#> 1        Emily   Patricia Balskus Summa Cum Laude female
#> 2        Aimee     Rose Candelore Summa Cum Laude female
```

```
#> 3        Megan   Elissa Delehanty Summa Cum Laude female
#> 4    Katherine  Keleher Desormeau Summa Cum Laude female
#> 5     Kristina        Gray Fisher Summa Cum Laude female
#> 6  Christopher      Edward Goggin Summa Cum Laude   male
#> 7      Johanna  Dorothy Heinrichs Summa Cum Laude female
#> 8      Bradley     Thomas Howells Summa Cum Laude   male
#> 9      Theresa Cunningham O'Brien Summa Cum Laude female
#> 10       Julia         Ann Snyder Summa Cum Laude female
#> 11        Adam  Hawthorne Steeves Summa Cum Laude   male
#> 12     Jessica        Ruth Bauman Magna Cum Laude female
#> 13       Laura      Marie Bennett Magna Cum Laude female
#> 14      Steven       James Biller Magna Cum Laude   male
#> 15       Laura Elizabeth Bothwell Magna Cum Laude female
```

'all.ratio' dataset is a table of the proportions of each gender from the class of 2003 to 2016. To get the dataset,

```
data(all.ratio)
all.ratio
```

```
#>    classyear    female      male
#> 1       2003 0.3903509 0.3122530
#> 2       2004 0.3476395 0.3692946
#> 3       2005 0.4273859 0.3117871
#> 4       2006 0.4000000 0.3117409
#> 5       2007 0.4039216 0.3183857
#> 6       2008 0.3755102 0.3153527
#> 7       2009 0.3815261 0.3144105
#> 8       2010 0.3703704 0.3333333
#> 9       2011 0.4440000 0.2703863
#> 10      2012 0.3636364 0.3501946
#> 11      2013 0.3729508 0.3348624
#> 12      2014 0.3166023 0.4140969
#> 13      2015 0.4078431 0.3276596
#> 14      2016 0.3568465 0.3628692
```

stat_rep()

All information can be presented in five statistical prepresentations using stat_rep(). Options are a summary table for each class year, the proportions of each gender over time, a summary of the proportions, a box plot of the proportions, and a hypothesis test[1].

```
stat_rep("annual")
```

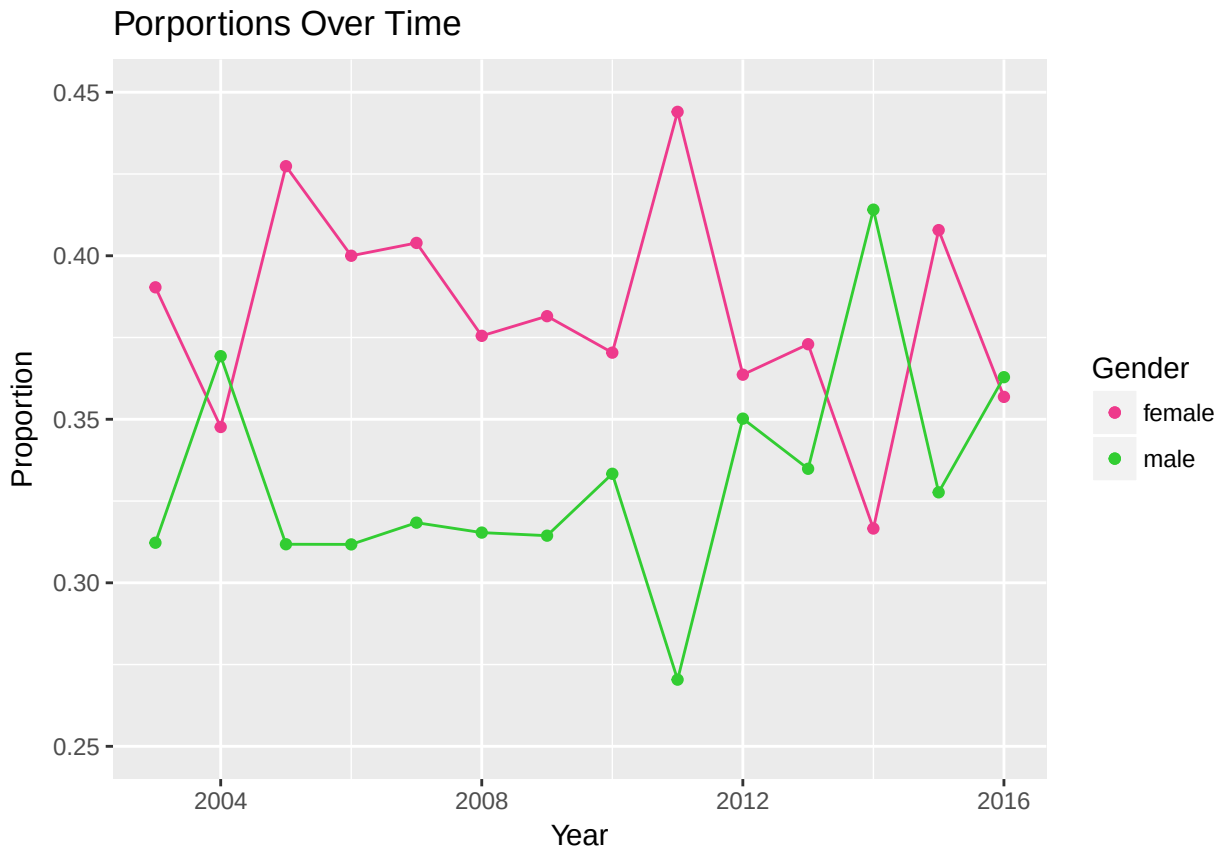
```
#> [[1]]
#>                   class of 2003 gender
#> honor              female male
#>   Summa Cum Laude      8    3
#>   Magna Cum Laude     30   33
#>   Cum Laude           51   43
#>   Sum                 89   79
#>
#> [[2]]
#>                   class of 2004 gender
#> honor              female male
#>   Summa Cum Laude      4    7
#>   Magna Cum Laude     31   33
#>   Cum Laude           46   49
#>   Sum                 81   89
#>
#> [[3]]
#>                   class of 2005 gender
#> honor              female male
```

[1] one sided, 95% confidence interval with alpha of 0.05

```
#>   Summa Cum Laude      3    7
#>   Magna Cum Laude     33   35
#>   Cum Laude           67   40
#>   Sum                103   82
#>
#> [[4]]
#>                  class of 2006 gender
#> honor             female male
#>   Summa Cum Laude      3    5
#>   Magna Cum Laude     34   30
#>   Cum Laude           57   42
#>   Sum                 94   77
#>
#> [[5]]
#>                  class of 2007 gender
#> honor             female male
#>   Summa Cum Laude      6    3
#>   Magna Cum Laude     34   28
#>   Cum Laude           63   40
#>   Sum                103   71
#>
#> [[6]]
#>                  class of 2008 gender
#> honor             female male
#>   Summa Cum Laude      4    6
#>   Magna Cum Laude     32   30
#>   Cum Laude           56   40
#>   Sum                 92   76
#>
#> [[7]]
#>                  class of 2009 gender
#> honor             female male
#>   Summa Cum Laude      6    3
#>   Magna Cum Laude     36   28
#>   Cum Laude           53   41
#>   Sum                 95   72
#>
#> [[8]]
#>                  class of 2010 gender
#> honor             female male
#>   Summa Cum Laude      5    4
#>   Magna Cum Laude     32   27
#>   Cum Laude           53   44
#>   Sum                 90   75
#>
#> [[9]]
#>                  class of 2011 gender
#> honor             female male
#>   Summa Cum Laude      5    4
#>   Magna Cum Laude     39   26
#>   Cum Laude           67   33
#>   Sum                111   63
#>
#> [[10]]
#>                  class of 2012 gender
#> honor             female male
#>   Summa Cum Laude      2    6
#>   Magna Cum Laude     28   38
#>   Cum Laude           54   46
#>   Sum                 84   90
#>
#> [[11]]
#>                  class of 2013 gender
#> honor             female male
```

```
#>   Summa Cum Laude      4    5
#>   Magna Cum Laude     39   30
#>   Cum Laude           48   38
#>   Sum                 91   73
#>
#> [[12]]
#>                 class of 2014 gender
#> honor            female male
#>   Summa Cum Laude      4    5
#>   Magna Cum Laude     23   44
#>   Cum Laude           55   45
#>   Sum                 82   94
#>
#> [[13]]
#>                 class of 2015 gender
#> honor            female male
#>   Summa Cum Laude      5    6
#>   Magna Cum Laude     44   25
#>   Cum Laude           55   46
#>   Sum                104   77
#>
#> [[14]]
#>                 class of 2016 gender
#> honor            female male
#>   Summa Cum Laude      4    6
#>   Magna Cum Laude     31   33
#>   Cum Laude           51   47
#>   Sum                 86   86
```

```
stat_rep("timeplot")
```

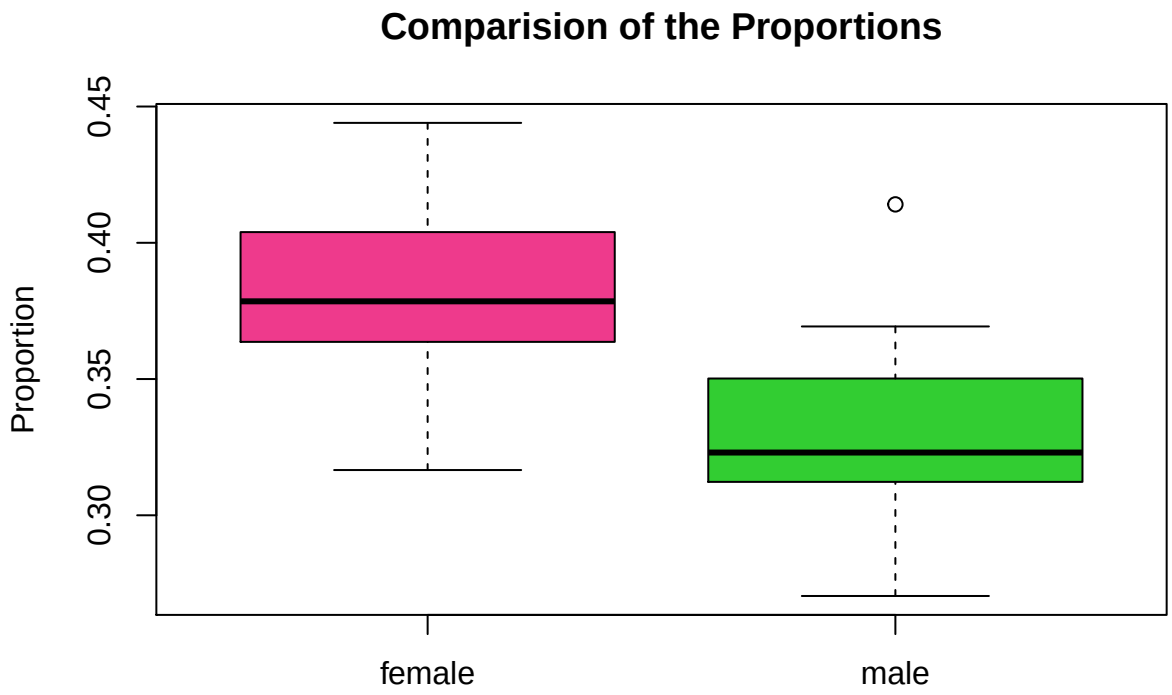



```
stat_rep("prop.sum")
```

```
#>         female   male
#> Min.    0.3166 0.2704
```

```
#> 1st Qu. 0.3653 0.3128
#> Median  0.3785 0.3230
#> Mean    0.3828 0.3319
#> 3rd Qu. 0.4029 0.3464
#> Max.    0.4440 0.4141
```

```
stat_rep("boxplot")
```




```
stat_rep("t.testing")
```

```
#>
#>  Welch Two Sample t-test
#>
#> data:  all.ratio$female and all.ratio$male
#> t = 4.0049, df = 25.964, p-value = 0.0002313
#> alternative hypothesis: true difference in means is greater than 0
#> 95 percent confidence interval:
#>  0.02919506        Inf
#> sample estimates:
#> mean of x mean of y
#> 0.3827560 0.3319019
```

Analysis

Exploiting the ready-to-use datasets in the package, the annual report of students of each gender who graduated with Latin honors are provided. Nevertheless, these figures take the total number of students for granted, causing a bias mentioned in the introduction section. Therefore, it is more reasonable to study the proportions of each gender than the absolute numbers.

The dataset '`all.ratio`' displays the proportions of students of each gender who graduated with Latin honors from the class of 2003 to 2016. Even though the total number of students of each gender are taken into account, the figures still implies difference gaps of gender proportions. The time plot allows a brief proportion comparison. Only in 2004, 2014, and 2016 were the male proportions above the female proportion. Essential statistical numbers are shown in five-summary table. All figures in female column are higher than the figures in the same row in male column. The difference is more obvious in the box plot when the whole box of female students is located higher than of male students. Despite these menifest observations, the difference gaps need to be proved whether it is significant.

A hypothesis test is conducted with the null hypothesis that the female proportion is not greater than the male proportion. With the p-value of 0.0001867, the result suggests the rejection of the null hypothesis. That is, the female proportion is significantly greater than the male proportion.

Conclusion

It can be concluded that female Williams students have been graduating with higher GPA (receving Latin honors) than male students have for over 10 years. The result suggests that the environments Williams College provides are more in favor of female students's achievement, in terms of GPA, than they are to male students.

Bibliography

L. Mullen. *gender: Predict Gender from Names Using Historical Data*, 2015. URL https://github.com/ropensci/gender. R package version 0.5.1. [p1]

Williams College Catalog Archive. Williams College, Williamstown, MA, USA, 2016. URL http://web.williams.edu/admin/registrar/catalog/archive.html. [p1]

Panchanok Jumrustanasan '19
Economics and Computer Science
Williams College
Williamstown, MA
pj4@williams.edu