Portfolio3

#### Dylan Vener

## Introduction

The department of Mathematics at Rose-Hulman Institute of Technology offers a wide variety of statistics classes, following a flat structure with minimal prerequisites. To this end, they created a class, Engineering Statistics 2 (MA 383), to serve as a terminating unit, for students who students who will not go on to take any other upper statistics classes.   
  
The department wishes to evaluate the performance of the class in this capacity, and devised a way of classifying how an upperlevel class was used by students; Bridging, the first class taken, before the rest of the curriculumn, Elective, the majority of classes, which can be taken in any order, and Terminal, which serves as the only upperlevel class taken by student, capping off their experience in the field.  
  
The second question they wanted answered was how the enrollment of the class has changed over time.

## Methodology

To solve this problem we were provided with the data for all students who have taken MA383 in the last 10 years, comprising a unique ID, all classes they took, whether they passed or failed, and the time the class was taken.   
  
To facilitate answering this questions, two distinct approaches were taken.   
  
To answer the first question, the dataframe was arranged by year, with a tie broked by term, a simple function was created to categorize each student's classes into the three categories provided by the department.

classify <- function(crs,ID){  
 #Returns a vector, wherein each position corresponds to Terminal, Bridging, Elective, and for error checking, N/A  
 if(!(ID %in% crs)){  
 return(c(0,0,0,1))  
 } else if(length(crs)==1){  
 return(c(1,0,0,0))  
 } else if(crs[1]==ID){  
 return(c(0,1,0,0))  
 } else{  
 return(c(0,0,1,0))  
 }  
}

Once I had this, I looped through the ID's, and counted up over the entire dataframe, to get the total. I did this for all classes, for a point of comparison.

#Create an empty array  
results.df <- tibble()  
#loops through all classes, initiates the sum to zero, and when done summing, puts it into a datafram  
for (k in classes){  
 results=c(0,0,0,0)  
 for (i in students){  
 results <- results+classify(filter(upper.df,ID==i)$CRSE,k)   
 }  
 rbind(results.df,tibble(course=k,term=results[1],branch=results[2],elec=results[3])) -> results.df  
}

To answer the second question, I used a builtin function to count up the instances of each class, so that I can show the popularity of the class in reference to other classes.

time <- as\_tibble(table(stats.df[,names(stats.df)%in% c("CRSE","YEAR")]))

## Conclusion

To best represent these two results, I created the following two graphs.

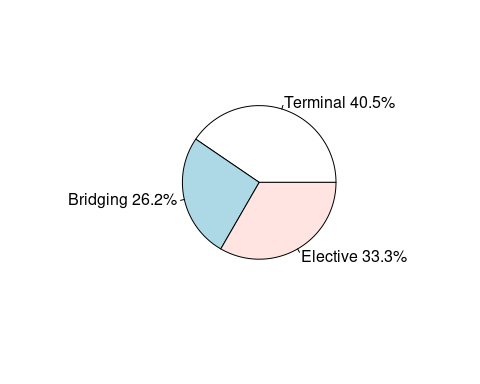
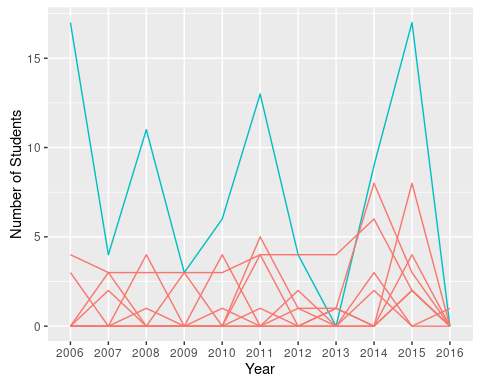
To answer the first question, I decided that a pie graph would best serve to represent the relationship between the three values. As well as labeling these, I decided to include the percentage of the total, to overcome the shortcomings of the pie chart, primarily loss of exact details. 

Fig.1-Pie chart of how the class is used.

As fig. 1 shows, the class serves its place as a terminator class, and further analysis shows that this is the only class in the curriculum to serve this purpose.

To answer the other question, I created a lineplot of all the classes, showing how the class's population as varied over the past 10 years, in comparison to all other classes.

ggplot(data=time,mapping=aes(x=YEAR,y=n,group=CRSE,color=CRSE=="383")) +  
 geom\_line() + xlab("Year") + ylab("Number of Students") +  
 theme(legend.position="none")

Fig. 2-in blue, MA383, everything else in red, number of students in the class each year.

As fig. 2 shows, Engineering statistics 2, is by far the most popular class most years for which its offered.

These two plots really show that the class has served its purpose well, it both acts as an upperlevel terminating class, and it's very popular, meaning that its getting a good coverage of the student population.

#To run on a different computer, dplyr, tibble, and ggplot2 must be installed, and they must have the dataset on their computer, and to modify the path to its location  
library('dplyr')  
library('tibble')  
library('ggplot2')  
ma.df <- as\_tibble(read.csv(file="~/Statistical Programming/MA383.csv",header=TRUE, stringsAsFactors=FALSE))  
ma.df <- ma.df%>%arrange(YEAR,TERM)  
#Filters to only upper level statistics classes  
upper.df <- ma.df %>% filter(CRSE %in% c(383,385,386,387,480,481,482,485,487))  
students <- unique(upper.df$ID)  
classes <- unique(upper.df$CRSE)  
results.df <- tibble()  
  
#counts up the instances of each class, for each year, and then plots each value  
time <- as\_tibble(table(upper.df[,names(upper.df)%in% c("CRSE","YEAR")]))  
ggplot(data=time,mapping=aes(x=YEAR,y=n,group=CRSE,color=CRSE=="383")) +  
 geom\_line() + xlab("Year") + ylab("Number of Students") + ggtitle('MA383 vs Other Upper Level Classes')  
 theme(legend.position="none")  
  
#Classifies a student's list of classes, into N/A, terminating, branching, or elective, for a class ID  
classify <- function(crs,ID){  
 if(!(ID %in% crs)){  
 return(c(0,0,0,1))  
 } else if(length(crs)==1){  
 return(c(1,0,0,0))  
 } else if(crs[1]==ID){  
 return(c(0,1,0,0))  
 } else{  
 return(c(0,0,1,0))  
 }  
}  
  
#Loops through each students, and counts up each classifications  
for (k in classes){  
 results=c(0,0,0,0)  
 for (i in students){  
 results <- results+classify(filter(upper.df,ID==i)$CRSE,k)   
 }  
 rbind(results.df,tibble(course=k,term=results[1],branch=results[2],elec=results[3])) -> results.df  
}  
  
#Because of certain difficulties, I resorted to harcoding the pie chart to the correct values, to ensure the clarity  
pie(c(34,22,28),labels=c('Terminal 40.5%','Bridging 26.2%','Elective 33.3%'))