Statistical Methods – COSC 6323 - Project-Milestone 1

By Dinesh Narlakanti (2083649) and Adarsh Chagantipati (2054031)

INTRODUCTION

A national wide survey of U.S.academics was conducted and around 400 intellectuals from 70 phD granting institutions, who excelled in the competitive grant process participated and filled a Core Questionnaire (CQ) which includes questions on their behavioral characteristics, proposal tactics, time they spend on research, sleep and other important things they do regularly. This survey and analysis of it concludes that long research hours and thoughtful choices are the primary reason for the grantmanhip and academic fame is the secondary reason for grantsmanship.

This document includes code development, analytic observations and visualization of gender distribution, disciplinary distribution, geographic distribution, weekly workload of the participants, research load of the participants and funding coverage for their proposals.

Familiarizing with the dataset and information about the columns used:

Gender Distribution
Columns used: Gender
Column description: Gender of the participants [Female, Male]

```
levels(factor(all_data$Gender))
```

```
## [1] "Female" "Male"
```

Disciplinary Distribution
Columns used: Department, Department_Other
Columns description: Department of the participants.

```
all_departments = c(all_data$Department, all_data$Department_Other)
levels(factor(all_departments[all_departments != 'Other']))
```

```
##  [1] ""
##  [2] "Applied Math"
##  [3] "Biology"
##  [4] "Chemistry"
##  [5] "Chemistry and Biochemistry"
##  [6] "Communication Studies"
##  [7] "Computer Science"
##  [8] "Engineering"
##  [9] "Engineering / Learning Sciences"
## [10] "Environmental Studies"
## [11] "Geosciences"
## [12] "Health"
## [13] "Health and Human Performance"
## [14] "HHP"
## [15] "Informatics"
## [16] "Information"
## [17] "Information and Logistics Technology"
## [18] "Information Science"
## [19] "Management"
## [20] "Mathematics"
## [21] "Medicine"
## [22] "Neurobiology"
## [23] "Neuroscience"
## [24] "Oceanography"
## [25] "Optometry"
## [26] "Pharmacological and Pharmaceutical Sciences"
## [27] "Pharmacy"
## [28] "Pharmacy "
## [29] "Physics"
## [30] "Political Science"
## [31] "Political Science "
## [32] "Psychology"
## [33] "Scientific Computing"
## [34] "Sociology"
## [35] "Speech, Language, and Hearing"
## [36] "Statistical Sciences and Operations Research"
## [37] "Technology"
## [38] "Vision Science"
```

Geographic Distribution

Columns used: State
Column description: States where the participants live.

```
levels(factor(all_data$State))
```

```
##  [1] "Alabama"        "Arizona"        "California"     "Colorado"      
##  [5] "Connecticut"    "Florida"        "Georgia"        "Illinois"      
##  [9] "Maryland"       "Massachusetts"  "Michigan"       "Minnesota"     
## [13] "Nevada"         "New Jersey"     "New Mexico"     "New York"      
## [17] "North Carolina" "Ohio"           "Oklahoma"       "Oregon"        
## [21] "Pennsylvania"   "Texas"          "Utah"           "Virginia"      
## [25] "Wisconsin"
```

Weekly workload

Columns used: WH
Column description: Working hours in a typical week including research, teaching, administration, and service/ outreach

```
levels(factor(all_data$WH))
```

```
## [1] "< 30"  "> 50"  "30-40" "40-50"
```

Research workload

Columns used: TWR Column description: Percentage of research in a typical week in 10 % incremental order

```
levels(factor(all_data$TWR))
```

```
##  [1] "0"   "10"  "20"  "30"  "40"  "50"  "60"  "70"  "80"  "90"  "100"
```

Funding coverage

Columns used: Research_Funded_By_External_Grants
Column description: Percentage of research operations funded by external grants

```
levels(factor(all_data$Research_Funded_By_External_Grants))
```

```
## [1] "100-75%"      "25-1%"        "50-25%"       "75-50%"       "Fully funded"
## [6] "Not funded"
```

Life Style Disruptions

Columns used: P_Disrupted_Research, P_Disrupted_Sleep, P_Disrupted_Diet, P_Disrupted_PA, P_Disrupted_IR.

Column descriptions: Relationships disruption introduced by proposal deadlines in a scale of [1-5].

Steps involved in Figures generation.

Step-1: For reproducing Disciplinary Distribution and Geographic Distribution, created an extra column named ‘dept’ to group few departments as one discipline and created another extra column named ‘geo’ to group states as regions. And for Research Load, we created an extra column named ‘rl’ to group the reseach workload.

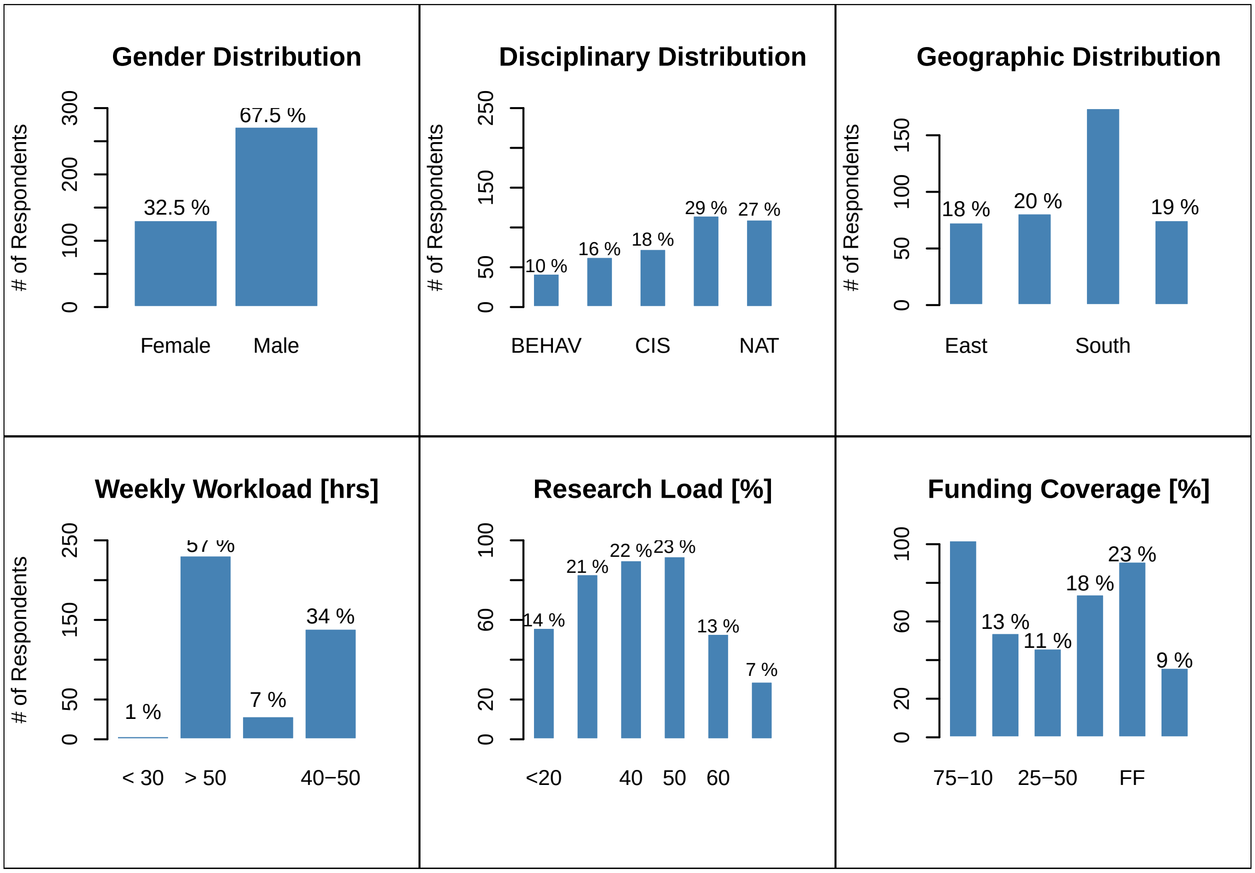
```
NAT <- c('Applied Math', 'Chemistry and Biochemistry', 'Oceanography',
         'Environmental Studies', 'Chemistry', 'Physics', 'Geosciences',
         'Mathematics')
CIS <- c( 'Informatics', 'Information', 'Information and Logistics Technology',
          'Information Science', 'Computer Science', 'Scientific Computing')
ENG <- c('Engineering / Learning Sciences', 'Management',
         'Statistical Sciences and Operations Research', 'Technology',
         'Engineering')
BIO <- c('Health', 'HHP','Health and Human Performance', 'Neurobiology',
         'Neuroscience', 'Optometry',
         'Pharmacological and Pharmaceutical Sciences', 'Pharmacy',
         'Vision Science', 'Biology', 'Medicine')
BEHAV <- c('Political Science', 'Sociology', 'Psychology',
           'Speech, Language, and Hearing', 'Political Science',
           'Communication Studies')

East <- c('Alabama', 'New Jersey', 'Massachusetts', 'Connecticut', 'New York',
          'Pennsylvania', 'North Carolina', 'Virginia',  'Maryland')
West <- c('Colorado', 'Arizona', 'California', 'Nevada', 'New Mexico', 'Oregon',
          'Utah')
Midwest <- c('Illinois', 'Michigan', 'Minnesota', 'Ohio', 'Wisconsin' )
South <- c('Florida', 'Georgia', 'Oklahoma', 'Texas')

all_data$dept <- ifelse(all_data$Department %in% NAT | all_data$Department_Other
                        %in% NAT, 'NAT',
                        ifelse(all_data$Department %in% CIS |
                                 all_data$Department_Other %in% CIS, 'CIS',
                               ifelse(all_data$Department %in% ENG |
                                        all_data$Department_Other %in% ENG,
                                      'ENG',
                                      ifelse(all_data$Department %in% BIO |
                                               all_data$Department_Other
                                             %in% BIO, 'BIO', 'BEHAV'))))

all_data$geo <- ifelse(all_data$State %in% East, 'East',
                       ifelse(all_data$State %in% West, 'West',
                              ifelse(all_data$State %in% Midwest, 'Midwest','South')))


all_data$rl <- ifelse(all_data$TWR <=20, 20,
                      ifelse(all_data$TWR == 30, 30,
                             ifelse(all_data$TWR == 40, 40,
                                    ifelse(all_data$TWR == 50, 50,
                                           ifelse(all_data$TWR == 60, 60, 70
                                                  )))))
```

Step-2: To display percentages as labels above the bars, created a variable to store percentages based on their repetitions.

```
n <- nrow(all_data)

gender_percentage <- table(all_data$Gender)/n * 100
dept_percentage <- table(all_data$dept)/n * 100
geo_percentage <- table(all_data$geo)/n * 100
wh_percentage <- table(all_data$WH)/n * 100
rl_percentage <- table(all_data$rl)/n * 100
fc_percentage <- table(all_data$Research_Funded_By_External_Grants)/n * 100
```

Step-3: Plotting the graphs using barplots and text() to represent percentage labels.

```
par(mfrow=c(2,3))
gender_dist <- barplot(table(all_data$Gender), ylim = c(0,300), xlim=c(0,2),
                       ylab = "# of Respondents", main="Gender Distribution",
                       col = "steelblue", border="white", width = 0.7)
gpf <- paste(round(gender_percentage[1],1),'%')
gpm <- paste(round(gender_percentage[2],1),'%')
text(0.5,150,gpf)
text(1.3,290,gpm)
box("figure")


dept_dist <- barplot(table(all_data$dept),ylab = "# of Respondents",
                     ylim = c(0,250), main="Disciplinary Distribution",
                     col = "steelblue",space = 1, border="white")
text(1.5,52,paste(round(dept_percentage[1]),'%'), cex = 0.9)
text(3.5,73,paste(round(dept_percentage[2]),'%'), cex = 0.9)
text(5.5,85,paste(round(dept_percentage[3]),'%'), cex = 0.9)
text(7.5,125,paste(round(dept_percentage[4]),'%'), cex = 0.85)
text(9.5,123,paste(round(dept_percentage[5]),'%'), cex = 0.9)
box("figure")

geo_dist <- barplot(table(all_data$geo),ylab="# of Respondents",
                    main="Geographic Distribution", col = "steelblue",
                    space =1,border="white")
text(1.5,85,paste(round(geo_percentage[1]),'%'))
text(3.6,92,paste(round(geo_percentage[2]),'%'))
text(5.6, 185,paste(round(geo_percentage[3]),'%'))
text(7.6,87,paste(round(geo_percentage[4]),'%'))
box("figure")


week_work <- barplot(table(all_data$WH),ylim= c(0,250),
                     ylab="# of Respondents", main="Weekly Workload [hrs]",
                     col = "steelblue", border="white")
text(0.7,35,paste(round(wh_percentage[1]),'%'))
text(2,245,paste(round(wh_percentage[2]),'%'))
text(3.1,50,paste(round(wh_percentage[3]),'%'))
text(4.3,155,paste(round(wh_percentage[4]),'%'))
box("figure")

research_load <- barplot(table(all_data$rl),ylim= c(0,100),
                         main="Research Load [%]", col = "steelblue",
                         border="white", space = 1,names.arg =
                           c('<20','30','40','50','60','>70'))
text(1.5,60, paste(round(rl_percentage)[1],'%'), cex = 0.9)
text(3.5,87, paste(round(rl_percentage)[2],'%'), cex = 0.9)
text(5.5,95, paste(round(rl_percentage)[3],'%'), cex = 0.9)
text(7.5,97, paste(round(rl_percentage)[4],'%'), cex = 0.9)
text(9.5,57, paste(round(rl_percentage)[5],'%'), cex = 0.9)
text(11.5,35, paste(round(rl_percentage)[6],'%'), cex = 0.9)
box("figure")

fund_cov <- barplot(table(all_data$Research_Funded_By_External_Grants),
                    main = "Funding Coverage [%]", border="white",
                    col = "steelblue", space =0.5, names.arg =
                      c('75-10','1-25','25-50','50-75','FF','NF'))
text(2.5,60,paste(round(fc_percentage)[2],'%'))
text(4,50,paste(round(fc_percentage)[3],'%'))
text(5.5,80,paste(round(fc_percentage)[4],'%'))
text(7,95,paste(round(fc_percentage)[5],'%'))
text(8.5,40,paste(round(fc_percentage)[6],'%'))
box("figure")
```

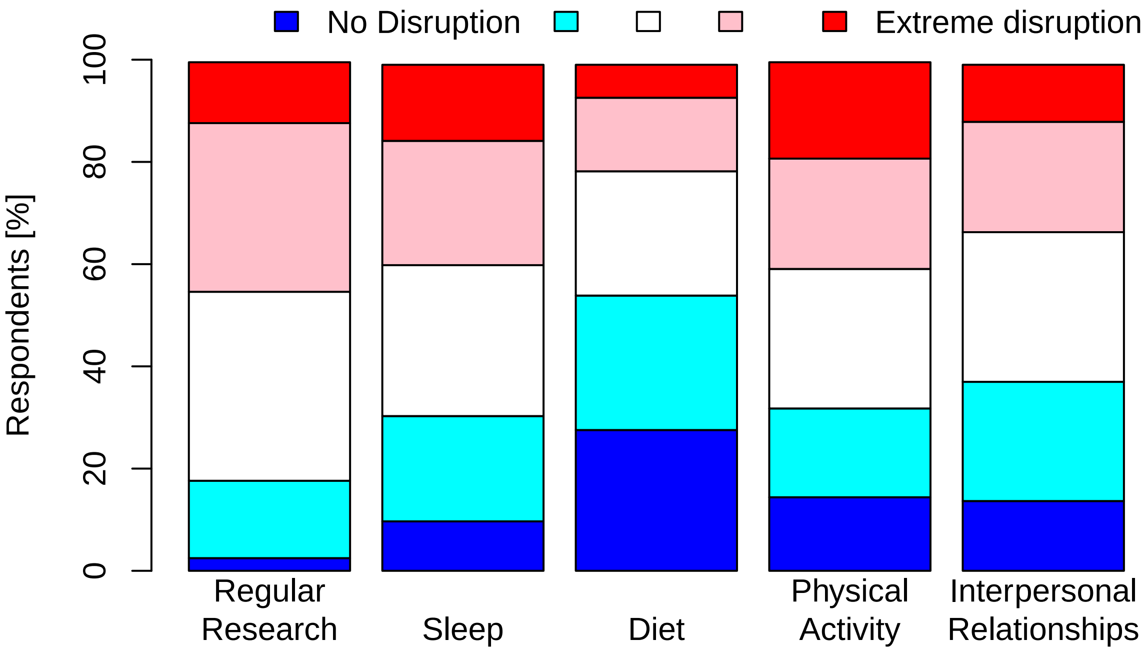


Step-4: Filtered rows from 44 to 48 that consists of life style disruptions. Added levels as (1,2,3,4,5). Using sapply created a table that calculates all the levels.

```
filtered <- all_data[44:48]
levs <- c(1,2,3,4,5)
table <- sapply(filtered, function(all_data) table(factor(all_data,
                                                          levels = levs, ordered = TRUE)))
```

Step-5: Plotted the stacked bar plot.

```
barplot(table/n*100,
        ylab = 'Respondents [%]',
        col = c("blue", "cyan",  "white","pink","red"), ylim=c(0,100),
        ,names.arg = c('Regular\nResearch', 'Sleep', 'Diet',
                       'Physical\nActivity', 'Interpersonal\nRelationships'),
        border = "black")
legend('top',legend = c("No Disruption","","","", 'Extreme disruption'),
       xpd =  TRUE, fill = c("blue", "cyan",  "white","pink","red"),
       inset = c(0,-0.15), bty ='n', horiz = TRUE,
       text.width =  c(0.50,0.01,0.01,0.01,0.10) )
```



ANALYTICAL OBSERVATIONS

1) Most of the Professors(participants) in the survey are male. Male professors are nearly doubled the number than female professors.
2) Engineering departments has the highest number of professors who are into research and applying for competitive grants.
3) Out of all the regions, Almost half (43%) of the professors are from south region. Number of professors living in East, Midwest and North are almost equal with a difference of 1%.
4) Majority of the professors spend more than 50 hours weekly on their academic related activities like

researching, teaching and service/ outreach. Very few professors, 1% of all spend less than 30 hours of their weekly time on academic related activities.
5) Majority of the professors who are working and applying for grants spends 50 hours of their weekly time purely on research. This clearly shows that the more amount the out on academic related activities, the higher the chances of getting grants for their research. This also shows that very few of professors, around 7% spend more than 70 hours on research.
6) Quarter percentage of the professors are granted and financially funded with 75-100% of their research. And only 10% of the total applications are not funded. This shows that 90% of the projects that are applied for the grants are funded.
7) Diet doesn't have any disruption because of the proposal deadline.
8) Due to proposal deadline, we can clearly see there is extreme disruption in physical activity first and second is the sleep.
9) However, there is no disruption in the regular research due to proposal deadlines.