Survey Data Analysis report

SHUANG SONG

2022-03-02

### 1. How did you prepare the data for analysis?

* First, I loaded the data using R language and did some data exploratory analysis. There are 668 rows and 125 columns.
* Second, reformat the column names. In the raw data there are question mark, dash, space, dot that is not suitable and standard for column names, so I filter them out and replaced them with either underscore, or no space.
* From instruction on power point, we need to exclude the international student, so I filtered and excluded them from the data frame.
* Now the data is much more clean and ready for further analysis. See code below.

load the packages we need:

rm(list = ls()) # clear the environment  
library(readxl)  
library(writexl)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyr)  
library(ggplot2)  
library(scales)  
library(ggpubr)

read in xlsx file :

file<- read\_excel("/Users/cleopathy/Desktop/assignment1.xlsx", sheet = "RAW DATA - DEIDENTIFIED")

## New names:  
## \* respect -> respect...9  
## \* intclimate -> intclimate...10  
## \* socclimate -> socclimate...11  
## \* collegial -> collegial...12  
## \* sameuniv -> sameuniv...13  
## \* ...

excluded international student:

file <- filter(file, Q62 != "International (Non-U.S. Citizen with temporary U.S. Visa)")

reformat the column names:

#reformat column name  
colnames(file) <- gsub("/", "\_", colnames(file))  
colnames(file) <- gsub(" ", "\_", colnames(file))  
colnames(file) <- gsub("\\?", "", colnames(file))  
colnames(file) <- gsub("\\.", "", colnames(file))  
head(file)

## # A tibble: 6 × 125  
## ResponseId Q6 Q8 `STEM\_Non-STEM` Q12 satacad satlife satoverall  
## <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 R\_2cCN179YVYmv5Cl PhD Coll… Non-STEM <NA> Very g… Good Good   
## 2 R\_6QjQtKUFNbhLsoV PhD Coll… Non-STEM <NA> Excell… Good Excellent   
## 3 R\_UEkRKRes2mmR5bX PhD Coll… STEM <NA> Very g… Good Good   
## 4 R\_3LgLwQCm1Il114u PhD Coll… Non-STEM I wi… Very g… Good Very good   
## 5 R\_3VmQbgty78CG8Df PhD Coll… STEM I wi… Very g… Very g… Very good   
## 6 R\_3QMoPUV7Zwm16na PhD Coll… STEM I wi… Fair Fair Fair   
## # … with 117 more variables: respect9 <chr>, intclimate10 <chr>,  
## # socclimate11 <chr>, collegial12 <chr>, sameuniv13 <chr>, samefield14 <chr>,  
## # recommend15 <chr>, curriculum16 <chr>, teaching17 <chr>, advising18 <chr>,  
## # candidacy19 <chr>, interdisc20 <chr>, employment21 <chr>, progqual22 <chr>,  
## # Q21 <chr>, Q22 <chr>, Q23 <chr>, Q24 <chr>, finance <chr>, infotech <chr>,  
## # space <chr>, library <chr>, lab <chr>, htopicadv <chr>, hresearchadv <chr>,  
## # hwritingadv <chr>, hacadadv <chr>, hnonacadadv <chr>, hemployadv <chr>, …

### 2. How did you address any missing values?

First, I did analysis and find out if there are cases that are completed empty (full of NAs)in the entire row, these records should be excluded for the further analysis. There are 0 rows has been removed for this step.

Second, find out how many NAs in each column, if the NA rate is higher than a cut out rate (eg. >50%) in the column, that columns will be excluded. There are 25 column has been removed for this step.

For the rest of table still has some NAs, I did the imputation to replace NAs into a value. Currently, i imputed using the mean of the column. Imputation is especially important in data analysis and there are lots of methods for data imputation, however, for this analysis I used averaged imputation. In this way, missing values are taken care of.

### 3. Which statistical methods did you use for your data analysis, and why?

When compare if there is a significant difference between URG students’ group and non-URG students’ group for the categories, I used pairwise t-test to calculate p value. If P value > 0.05, there is no significant statistical difference between URG and Non-URG students, otherwise p value < 0.05, significant difference between these two groups.

For 3 groups or more than 3 groups comparison, eg. in ethnicity/race group, I used ANOVA to conduct an analysis showing whether there is significant difference of extent of helpful among these categories.

### 4. Did you determine response rates for the different questions? Why or why not?

Yes. In different analysis we have to consider and determine the response rate as they provide valuable insight of the accuracy of the data. The higher response rate in the data means it is more representative of samples, which is the purpose of the survey. In different questions, the response rate is the ratio of number of participants in study to the number of participants that were asked to join in the survey.

### 5. Did you consider weighting any of the data? If yes, why?

Yes. For example, from this data we can see in race/ethnicity variable, the percentage of white is much more than other group, making sample result biased toward white. This might be the issue of method of sampling, or they have higher response than other group. What we need to do is to have the target proportion, divided by the actual proportion of different race group, to get weight of each different group. In this way, it can be a re-balance way and can make survey more accurately to reflect population.

### 6. What appropriate tests of statistical significance did you consider using (and used), when evaluating differences across different sub-populations?

ANOVA is the method to use evaluating differences across different sub-population group. The ANOVA’s null hypothesis is that there is no difference in means of different groups, while the alternate hypothesis is the means are not all equal. There is a package of ANOVA in R to do this test. And if the p value is great than 0.05, then we will conclude that there is no difference in means of different groups. Otherwise, we will come to the conclusion that the means are not all equal.