Data 606 - Final Project

Avery Davidowitz

2022-12-08

Abstract

Market research is a data and time intensive field where it is often difficult to determine effectiveness. The goal of this analysis is to identify potential areas of interest for further analysis. Email marketing data will be subset to determine if there are statistically significant differences in effectiveness by the recipient's seniority. Therefore, the data set is observational with the observations being specific emails sent. It has a sample size of 1449. The analysis method used will be a chi-squared test for the proportions with a p-value alpha of .1. The findings will dictate if further A/B testing is appropriate to try to improve specific seniority subsets or if the campaigns should be targeted to specific demographics. The analysis does indicate that there is not enough evidence to support any conclusion that seniority is a factor in opening or response rates.

Introduction

```
# load data
url <- "https://raw.githubusercontent.com/adavidowitz100/DATA606/main/FinalProject/email_data.csv"

raw_email_data <- readr::read_csv(url, col_types = list(
              Seniority = col_factor(),
              Function = col_factor(),
              "Subject Line" = col_character(),
              "Hook 1" = col_character(),
              "Z_Contact Status" = col_factor(),
              "Email Opened" = col_date(format = "%m/%d/%Y"),
              "Email Replied" = col_date(format = "%m/%d/%Y")
))

head(raw_email_data)
```

```
## # A tibble: 6 x 7
##   Seniority Function     `Subject Line`    Hook ~1 Z_Con~2 Email Op~3 Email Re~4
##   <fct>     <fct>        <chr>             <chr>   <fct>   <date>     <date>
## 1 Director  Logistics    "Your participat~ "Impre~ Unresp~ NA         NA
## 2 Director  Distribution "Your distributi~ "Looks~ Attemp~ NA         NA
## 3 VP        Supply Chain "Your employee v~ "Saw y~ Attemp~ NA         NA
## 4 Director  Logistics    "Your insights o~ "Enjoy~ Unresp~ NA         NA
## 5 VP        Supply Chain "You + FLEXE Hyb~ "Impre~ Unresp~ NA         NA
## 6 VP        Distribution "Your participat~ "Read ~ Unresp~ 2022-07-20 NA
## # ... with abbreviated variable names 1: `Hook 1`, 2: `Z_Contact Status`,
## #   3: `Email Opened`, 4: `Email Replied`
```

```
df <- raw_email_data |> dplyr::rename(seniority = Seniority,
                                      department = Function,
                                      subject = "Subject Line",
                                      hook = "Hook 1",
                                      status = "Z_Contact Status",
                                      opened_date = "Email Opened",
                                      replied_date = "Email Replied") |>
                        dplyr::mutate(opened = if_else(!is.na(opened_date), TRUE, FALSE)) |>
                        dplyr::mutate(replied = if_else(!is.na(replied_date), TRUE, FALSE)) |>
                        dplyr::mutate(days_diff = difftime(opened_date, replied_date, units = "days"))
df$seniority[df$seniority=="Senior"] <- "Sr. Manager"
df$seniority[df$seniority=="Vp"] <- "VP"
df$seniority <- droplevels(df$seniority)
head(df)
```

```
## # A tibble: 6 x 10
##   seniority department subject hook  status opened_d~1 replied_~2 opened replied
##   <fct>     <fct>      <chr>   <chr> <fct>  <date>     <date>     <lgl>  <lgl>
## 1 Director  Logistics  "Your ~ "Imp~ Unres~ NA         NA         FALSE  FALSE
## 2 Director  Distribut~ "Your ~ "Loo~ Attem~ NA         NA         FALSE  FALSE
## 3 VP        Supply Ch~ "Your ~ "Saw~ Attem~ NA         NA         FALSE  FALSE
## 4 Director  Logistics  "Your ~ "Enj~ Unres~ NA         NA         FALSE  FALSE
## 5 VP        Supply Ch~ "You +~ "Imp~ Unres~ NA         NA         FALSE  FALSE
## 6 VP        Distribut~ "Your ~ "Rea~ Unres~ 2022-07-20 NA         TRUE   FALSE
## # ... with 1 more variable: days_diff <drtn>, and abbreviated variable names
## #   1: opened_date, 2: replied_date
```

Summary Statistics

The below contingency tables and proportion tables describe the distribution of the data by seniority.

```
open_xtab <- xtabs(~ seniority + opened, data=df)
open_xtab
```

```
##              opened
## seniority     FALSE TRUE
##   Director      372  428
##   VP            171  190
##   Sr. Manager    57   53
##   Head           16   18
##   Manager        54   78
##   C suite         2    7
##   Partner         0    3
```

```
open_prop <- prop.table(open_xtab, 1)
open_prop
```

```
##              opened
## seniority         FALSE      TRUE
##   Director    0.4650000 0.5350000
##   VP          0.4736842 0.5263158
```

```
##   Sr. Manager 0.5181818 0.4818182
##   Head        0.4705882 0.5294118
##   Manager     0.4090909 0.5909091
##   C suite     0.2222222 0.7777778
##   Partner     0.0000000 1.0000000
```

```
replied_xtab <- xtabs(~ seniority + replied, data=df)
replied_xtab
```

```
##              replied
## seniority     FALSE TRUE
##   Director      749   51
##   VP            333   28
##   Sr. Manager    98   12
##   Head           34    0
##   Manager       122   10
##   C suite         8    1
##   Partner         3    0
```

```
replied_prop <- prop.table(replied_xtab, 1)
replied_prop
```

```
##              replied
## seniority          FALSE       TRUE
##   Director    0.93625000 0.06375000
##   VP          0.92243767 0.07756233
##   Sr. Manager 0.89090909 0.10909091
##   Head        1.00000000 0.00000000
##   Manager     0.92424242 0.07575758
##   C suite     0.88888889 0.11111111
##   Partner     1.00000000 0.00000000
```

Due to the conditions for the Chi Squared test, some seniority levels will have to be excluded due to small sample and response sizes. C suite and Partner do not have enough in the opened/false case and Head needs to be excluded for the replied/true case. Independence should be satisfied by the relative size of the study compared to the total population of supply chain and logistics professionals in the US (although it may be possible that two individuals known to each other may receive the emails this is unlikely enough to ignore).

```
condition_pass_df <- df |> dplyr::filter(!seniority %in% c("C suite", "Partner", "Head"))
condition_pass_df$seniority <- droplevels(condition_pass_df$seniority)

pass_opened_xtab <- xtabs(~ seniority + opened, data=condition_pass_df)
pass_opened_prop <- prop.table(pass_opened_xtab, 1)
pass_opened_prop
```

```
##              opened
## seniority         FALSE      TRUE
##   Director    0.4650000 0.5350000
##   VP          0.4736842 0.5263158
##   Sr. Manager 0.5181818 0.4818182
##   Manager     0.4090909 0.5909091
```

```
pass_replied_xtab <- xtabs(~ seniority + replied, data=condition_pass_df)
pass_replied_prop <- prop.table(pass_replied_xtab, 1)
pass_replied_prop
```

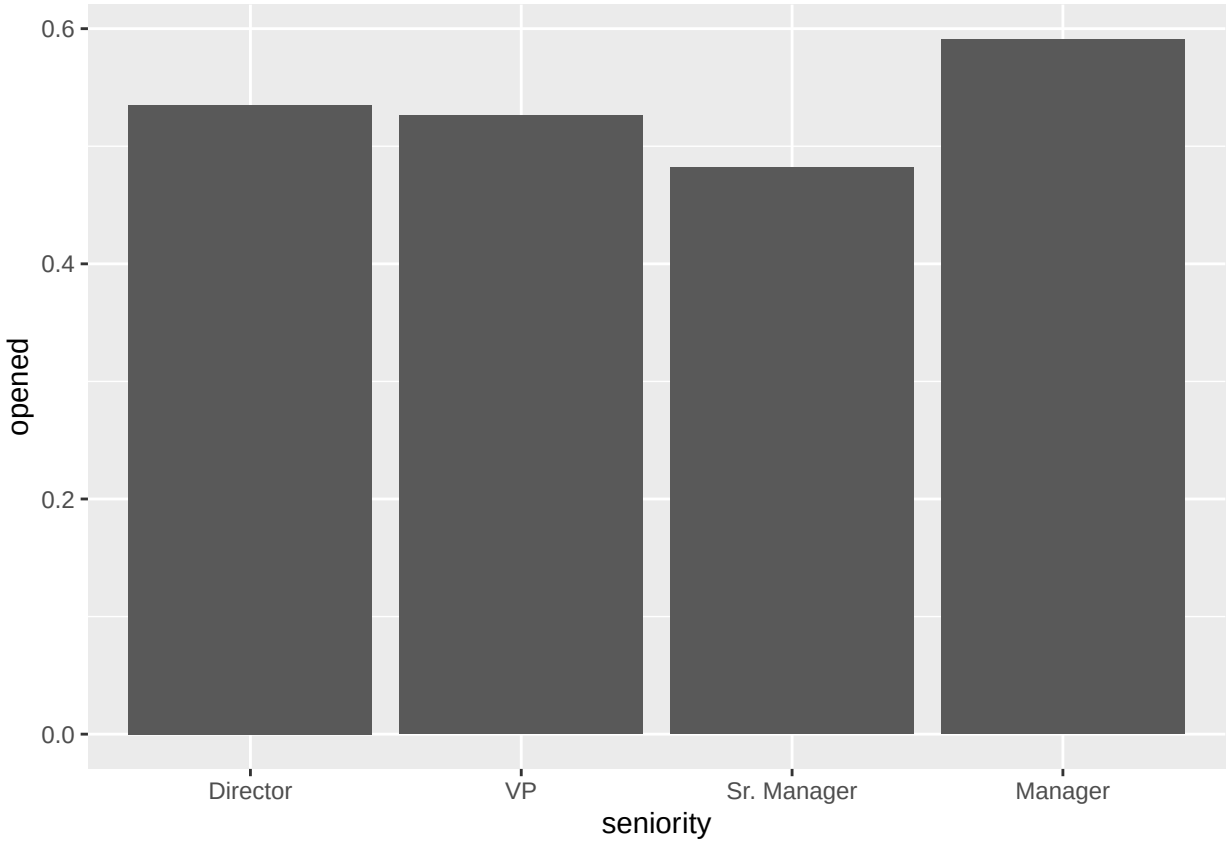
```
##              replied
## seniority          FALSE       TRUE
##   Director    0.93625000 0.06375000
##   VP          0.92243767 0.07756233
##   Sr. Manager 0.89090909 0.10909091
##   Manager     0.92424242 0.07575758
```

Data Visualizations

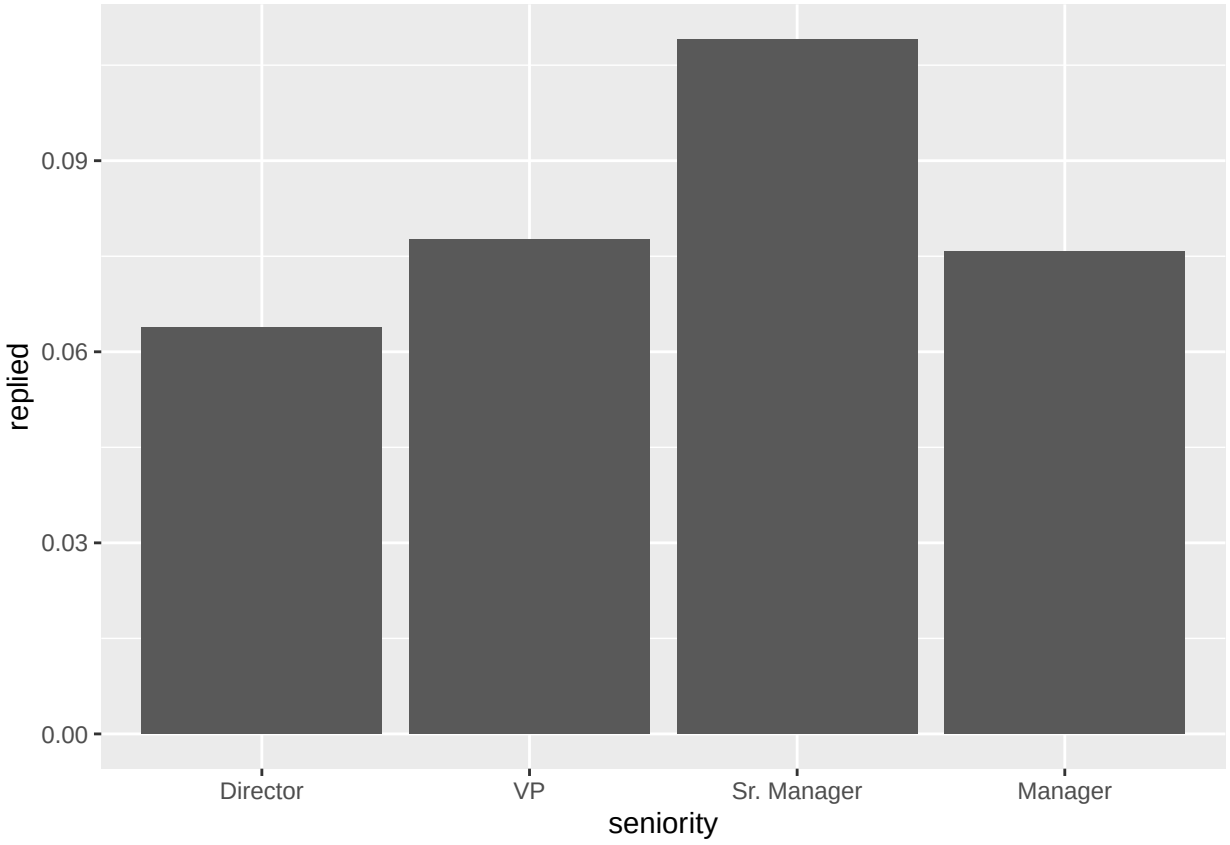
The below visualization seem to suggest that there is not much of a relationship between open rates and seniority. However, There does seem to be a more pronounced difference when looking at the differences at response rates.

```
percent_opened_df <- as.data.frame(pass_opened_prop)
percent_opened_df <- filter(percent_opened_df, opened==TRUE)
percent_replied_df <- as.data.frame(pass_replied_prop)
percent_replied_df <- filter(percent_replied_df, replied==TRUE)
percent_df <- left_join(percent_replied_df, percent_opened_df, by="seniority")
percent_df <- percent_df |> select(seniority, Freq.x, Freq.y) |> rename(replied=Freq.x, opened=Freq.y)

ggplot(data=percent_df, aes(x=seniority, y=opened)) +  geom_bar(stat="identity")
```

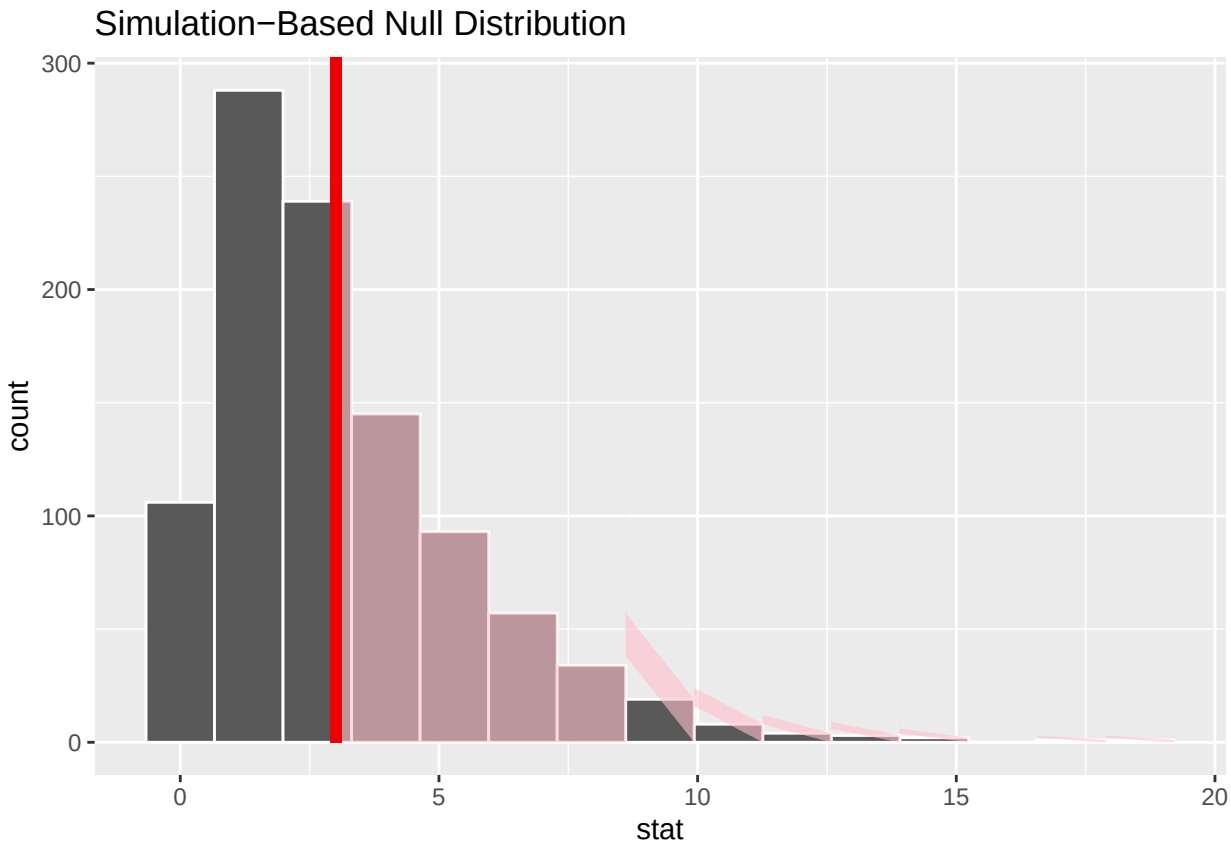


```
ggplot(data=percent_df, aes(x=seniority, y=replied)) +  geom_bar(stat="identity")
```



Statistical Tests

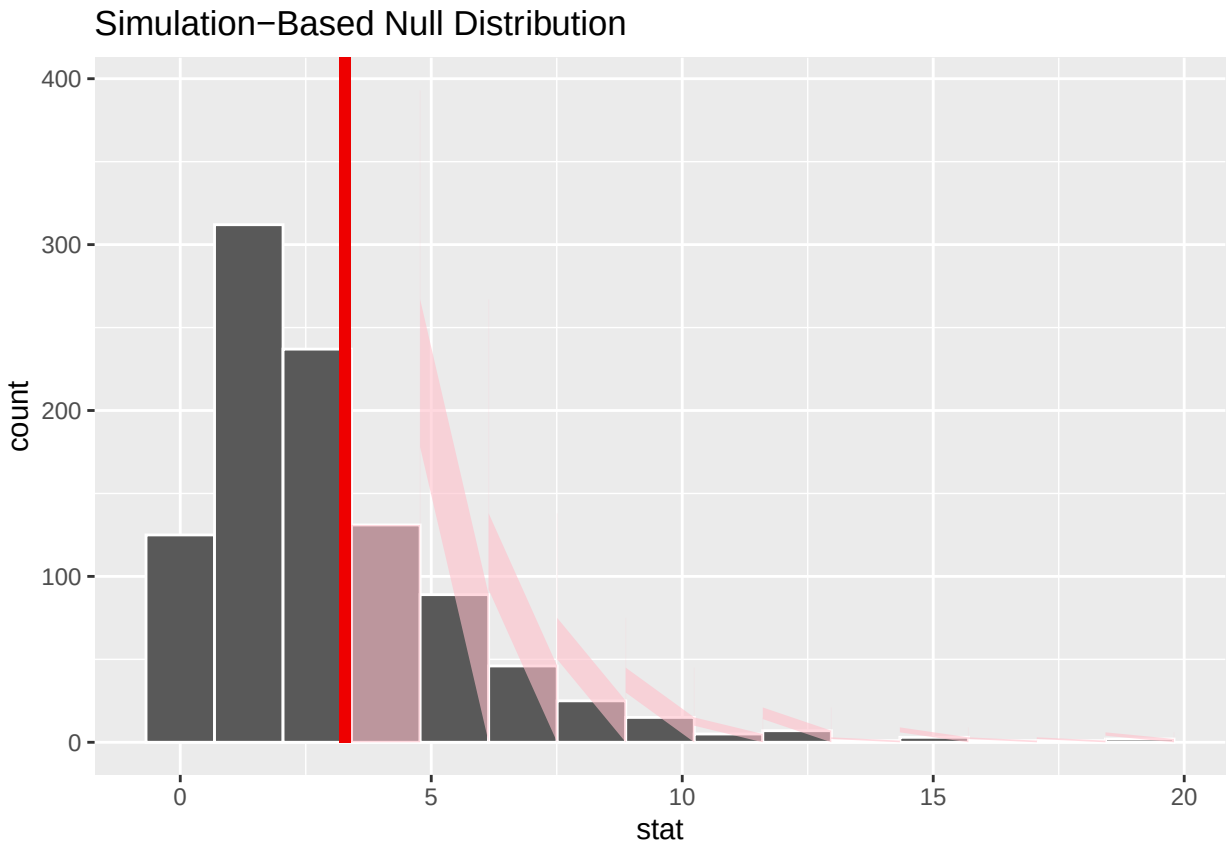
```
Chisq_hat <- condition_pass_df %>%
  specify(formula = opened ~ seniority) %>%
  hypothesize(null = "independence") %>%
  calculate(stat = "Chisq")
null_dist <- condition_pass_df %>%
  specify(opened ~ seniority) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "Chisq")
visualize(null_dist) +
  shade_p_value(obs_stat = Chisq_hat, direction = "greater")
```




```
null_dist %>%
  get_p_value(obs_stat = Chisq_hat, direction = "greater")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.421
```

```
Chisq_hat <- condition_pass_df %>%
  specify(formula = replied ~ seniority) %>%
  hypothesize(null = "independence") %>%
  calculate(stat = "Chisq")
null_dist <- condition_pass_df %>%
  specify(replied ~ seniority) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "Chisq")
visualize(null_dist) +
  shade_p_value(obs_stat = Chisq_hat, direction = "greater")
```




```
null_dist %>%
  get_p_value(obs_stat = Chisq_hat, direction = "greater")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.341
```

Conclusion

Since both of our p-values are very high and well over the alpha, we can conclude that there is not sufficient evidence to conclude that seniority and open and response rates are dependent. As a result, subsetting the marketing campaign by seniority does not appear to be a method to improve campaign performance. The fact that the one seniority level seemed to have a difference in response rate is quite possibly due to chance.