TT Assignment

David Syck

May 4, 2018

```
STAAR <- read.csv("./DATA/STAAR16-17.csv",
                  stringsAsFactors = FALSE,
                  na.strings = c(".","-1","-2","-3"))

STAAR <- subset(STAAR, GRDTYPE != "B")
STAAR$GRDTYPE <- factor(STAAR$GRDTYPE)
```

I load the data, treating periods and negative values as N/A. According to TEA these are masking symbols to protect the privacy of students. I also removed the "Both" designation from school types since only elementary, middle, and high school were requested.

```
#Should be 59976 obs
STAAR.gather <- STAAR %>%
  gather_(key = "headers", value = "value", names(STAAR)[-c(1,2)])

#Split headers into parts
STAAR.split<- within(STAAR.gather,{
  demographics <- substr(headers, 1,4)
  subject <- substr(headers,5,9)
  year <- substr(headers,10,11)
  n.d.r <- substr(headers,12,12)
})[-3]

#Save this to have a complete long format data set on hand
write.csv(STAAR.split,"./DATA/STAAR_Long.csv")



#Should be half the observations, 29988
STAAR.final <- STAAR.split[STAAR.split$year == 17,]

#I don't have a use for rate at the moment and it tidies things up a bit
#Should see 1/3 of the observations go away, 19992 left
STAAR.final <- subset(STAAR.final, n.d.r != "R")

STAAR.final[c(1,2,4,5,6,7)] <- lapply(STAAR.final[c(1,2,4,5,6,7)],factor)
#Every factor has the expected number of levels

#Rename demographics to something readable
STAAR.final$demographics <- plyr::revalue(STAAR.final$demographics,
               c(C200 = "two.races",
                 C300 = "asian",
                 C400 = "pacific.islander",
                 CA00 = "all",
                 CB00 = "african.american",
                 CE00 = "econ.dis",
                 CF00 = "female",
                 CH00 = "hispanic",
```

```
                CI00 = "american.indian",
                CL00 = "ell",
                CM00 = "male",
                CR00 = "risk",
                CS00 = "special.ed",
                CW00 = "white"))

#I prefer to have all the factors next to each other
STAAR.final <- STAAR.final[,c(1,2,4,5,6,7,3)]


#Check that summary of values is sensible
summary(STAAR.final$value)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0      31      87     151     179    3227    8109
```

```
#Remove data frames that I won't be using for the rest of the analysis
rm(STAAR, STAAR.gather, STAAR.split)
```

I think that in order to find percentages the data needs to go from wide to long and the information locked into the headers needs to be pulled out. I haven't worked with a lot of wide data so this was new for me. Some searching told me that gather() from the tidyr package was fairly standard.

The summary of the values is fine. Mostly I wanted to be sure that there were no negative numbers creeping in and that there wasn't some impossibly huge maximum. There are an awful lot of NA's which I should keep an eye on going forward.

```
#Expecting 1/3 of the observations, 6664
meets <- STAAR.final[STAAR.final$subject =="AR042",]
head(meets)
```

```
##         CAMPUS GRDTYPE n.d.r year subject demographics value
## 29989 57905001       S     N   17   AR042          all   434
## 29990 57905002       S     N   17   AR042          all   246
## 29991 57905003       S     N   17   AR042          all    86
## 29992 57905005       S     N   17   AR042          all   432
## 29993 57905006       S     N   17   AR042          all   164
## 29994 57905007       S     N   17   AR042          all   263
```

```
approaches <- STAAR.final[grep("AR01",STAAR.final$subject),]
head(approaches)
```

```
##         CAMPUS GRDTYPE n.d.r year subject demographics value
## 9997  57905001       S     D   17   AR010          all  1432
## 9998  57905002       S     D   17   AR010          all  1175
## 9999  57905003       S     D   17   AR010          all   226
## 10000 57905005       S     D   17   AR010          all  1513
## 10001 57905006       S     D   17   AR010          all   692
## 10002 57905007       S     D   17   AR010          all  1333
```

```
#Making racial and gender subset dataframes.
#I included all students in both of these subsets
meets.gender <- meets[meets$demographics %in% c("male",
                                                "female"),]

meets.race <- meets[meets$demographics %in% c("african.american",
```

```
                                                "white",
                                                "hispanic",
                                                "american.indian",
                                                "two.races",
                                                "asian",
                                                "pacific.islander"),]

approaches.gender <- approaches[approaches$demographics %in% c("male",
                                                               "female"),]

approaches.race <- approaches[approaches$demographics %in% c("african.american",
                                                "white",
                                                "hispanic",
                                                "american.indian",
                                                "two.races",
                                                "asian",
                                                "pacific.islander"),]
```

I split the data set into meets and approaches and then further into gender and race. This makes the downstream analysis more intuitive for me.

```
#I use dcast from the reshape2 package to aggregate
#And turn numerator/denominator into columns
overall.percents.m <-
  dcast(na.omit(meets[meets$demographics == "all",]),
        demographics~n.d.r, fun.aggregate = sum, value.var = "value") %>%
  mutate(percentage.m = N/D)

gradetype.percents.m <-
  dcast(na.omit(meets),
        GRDTYPE~n.d.r, fun.aggregate = sum, value.var = "value") %>%
  mutate(percentage.m = N/D)

race.percents.m <-
  dcast(na.omit(meets.race),
        demographics~n.d.r, fun.aggregate = sum, value.var = "value") %>%
  mutate(percentage.m = N/D)

gender.percents.m <-
  dcast(na.omit(meets.gender),
        demographics~n.d.r, fun.aggregate = sum, value.var = "value") %>%
  mutate(percentage.m = N/D)

overall.percents.a <-
  dcast(na.omit(approaches[approaches$demographics == "all",]),
        demographics~n.d.r, fun.aggregate = sum, value.var = "value") %>%
  mutate(percentage.m = N/D)

gradetype.percents.a <-
  dcast(na.omit(approaches), GRDTYPE~n.d.r, fun.aggregate = sum, value.var = "value") %>%
  mutate(percentage.a = N/D)

race.percents.a <-
  dcast(na.omit(approaches.race),
```

```
        demographics~n.d.r, fun.aggregate = sum, value.var = "value") %>%
  mutate(percentage.a = N/D)

gender.percents.a <-
  dcast(na.omit(approaches.gender),
        demographics~n.d.r, fun.aggregate = sum, value.var = "value") %>%
  mutate(percentage.a = N/D)

overall.final <-
  full_join(overall.percents.m[,c(1,4)],overall.percents.a[,c(1,4)], by = "demographics")
rm(overall.percents.m,overall.percents.a)

grade.final <-
  full_join(gradetype.percents.m[,c(1,4)],gradetype.percents.a[,c(1,4)], by = "GRDTYPE")
rm(gradetype.percents.m,gradetype.percents.a)

race.final <-
  full_join(race.percents.m[,c(1,4)],race.percents.a[,c(1,4)], by = "demographics")
rm(race.percents.m,race.percents.a)

gender.final <-
  full_join(gender.percents.m[,c(1,4)],gender.percents.a[,c(1,4)], by = "demographics")
rm(gender.percents.m,gender.percents.a)

kable(overall.final,
      col.names = c("", "Meets", "Approaching"),
      digits = 3)
```

	Meets	Approaching
all	0.382	0.63

```
kable(grade.final,
      col.names = c("School Type", "Meets", "Approaching"),
      digits = 3)
```

School Type	Meets	Approaching
E	0.388	0.671
M	0.280	0.618
S	0.403	0.511

```
kable(race.final,
      col.names = c("Race", "Meets", "Approaching"),
      digits = 3)
```

Race	Meets	Approaching
two.races	0.758	0.917
asian	0.667	0.658
african.american	0.289	0.528
hispanic	0.383	0.644
white	0.762	0.867

Race	Meets	Approaching
american.indian	NA	0.593

```
kable(gender.final,
      col.names = c("Gender", "Meets", "Approaching"),
      digits = 3)
```

Gender	Meets	Approaching
female	0.422	0.678
male	0.344	0.585

The four tables of percentages for submission. Finding a cleaner, less verbose way of doing this would be a high priority when revisiting the problem. I used kable to make nicer tables but I think there's a lot of room to make them prettier and, more importantly, more effective at conveying information. This code block is what I'm least thrilled about.

As far as how much sense the data makes:

All the values are between 0 and 1. This is good

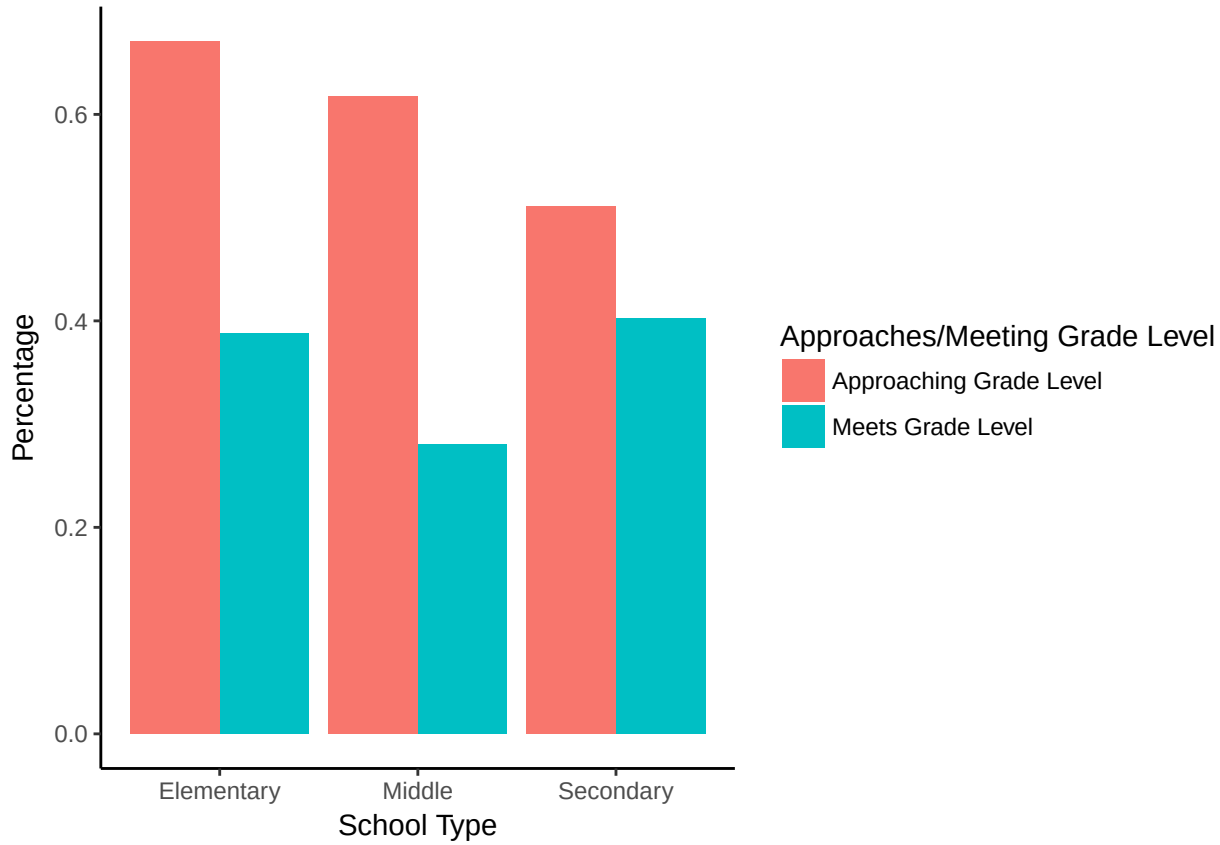
The total percentage of meets and approaches are over 1. This is likely because the approach data includes approaches or above.

If that's true then all approach percentages should be over the meets percentages and the Asian demographic is not. This is worrisome.

The American Indian demographic is NA in meets. Checking the meets.race data frame does show that there are no values for American Indians.

```
grade.final.long <- grade.final %>%
  gather(key = approach.meeting, value = percentage,names(grade.final)[-1])

percent.bar <-
  ggplot(grade.final.long, aes(x = GRDTYPE, y = percentage, fill = approach.meeting )) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_discrete(name = "Approaches/Meeting Grade Level",
                      labels = c("Approaching Grade Level", "Meets Grade Level")) +
  xlab("School Type") + ylab("Percentage") +
  scale_x_discrete(labels = c("Elementary", "Middle", "Secondary")) +
  theme_classic()
percent.bar
```




```
ggsave("SchoolTypeGradeLevels.jpg", plot = percent.bar)
```

```
## Saving 6.5 x 4.5 in image
```

```
rm(grade.final.long)
```

A bar chart seems like the obvious choice for showing the difference between variables.

Assumptions

While the documentation on variables regarding students meeting grade level was verbose and clear about what the numerator and denominator meant the approaches documentation was more complex and I had to make some assumptions.

1. The numerator column was the count of students that reached the specified grade level
2. The denominator column was the total count of students that took the test +In my exploration code I found that the denominators are different for the approaching and meets datasets. If this assumption is true that difference seems odd.