Statistical Sampling- Unit 1 - NAEP statistics

Alex Deshowitz

April 30, 2017

The dataframe used in this quick analysis was pulled from the NAEP nation's report card reports. The data was pulled at the state level for 2009 - 2015. The scores included are Math scores only.

More information regarding the data source can be found here : <https://nces.ed.gov/nationsreportcard/>

This very high level analysis covers the state of Texas and the changes in the scores for the state over the past 7 years.

packages <- c('plyr', 'sqldf', 'reshape','ggplot2')  
  
sapply(packages, library, character.only = TRUE)

## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

##   
## Attaching package: 'reshape'

## The following objects are masked from 'package:plyr':  
##   
## rename, round\_any

## $plyr  
## [1] "plyr" "stats" "graphics" "grDevices" "utils" "datasets"   
## [7] "methods" "base"   
##   
## $sqldf  
## [1] "sqldf" "RSQLite" "gsubfn" "proto" "plyr"   
## [6] "stats" "graphics" "grDevices" "utils" "datasets"   
## [11] "methods" "base"   
##   
## $reshape  
## [1] "reshape" "sqldf" "RSQLite" "gsubfn" "proto"   
## [6] "plyr" "stats" "graphics" "grDevices" "utils"   
## [11] "datasets" "methods" "base"   
##   
## $ggplot2  
## [1] "ggplot2" "reshape" "sqldf" "RSQLite" "gsubfn"   
## [6] "proto" "plyr" "stats" "graphics" "grDevices"  
## [11] "utils" "datasets" "methods" "base"

overall\_df <- read.csv('Analysis/Data/Math\_educ\_overall\_v01.csv', header = TRUE, skip = 9, nrows = 208)

head(overall\_df)

## Year Jurisdiction Average.scale.score Standard.Error  
## 1 2015 Alabama 267 1.2370776  
## 2 NA Alaska 280 0.9619621  
## 3 NA Arizona 283 1.3933827  
## 4 NA Arkansas 275 1.4291864  
## 5 NA California 275 1.3250740  
## 6 NA Colorado 286 1.4710870

ncol(overall\_df)

## [1] 4

str(overall\_df)

## 'data.frame': 208 obs. of 4 variables:  
## $ Year : int 2015 NA NA NA NA NA NA NA NA NA ...  
## $ Jurisdiction : Factor w/ 52 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ Average.scale.score: int 267 280 283 275 275 286 284 280 263 291 ...  
## $ Standard.Error : num 1.237 0.962 1.393 1.429 1.325 ...

summary(overall\_df)

## Year Jurisdiction Average.scale.score Standard.Error   
## Min. :2009 Alabama : 4 Min. :254 Min. :0.5051   
## 1st Qu.:2010 Alaska : 4 1st Qu.:278 1st Qu.:0.8355   
## Median :2012 Arizona : 4 Median :284 Median :1.0131   
## Mean :2012 Arkansas : 4 Mean :283 Mean :1.0142   
## 3rd Qu.:2014 California: 4 3rd Qu.:288 3rd Qu.:1.2044   
## Max. :2015 Colorado : 4 Max. :301 Max. :1.7877   
## NA's :204 (Other) :184

The first thing we need to do is create a better year column

years <- unique(overall\_df$Year[!is.na(overall\_df$Year)])  
  
y <- rep(years, each = 52)  
  
overall\_df['new\_year'] <- y  
  
# drop the old column  
  
overall\_df$Year <- NULL

Let's reorder the dataframe now

overall\_df<- overall\_df[,c("new\_year","Jurisdiction","Average.scale.score","Standard.Error")]  
  
# rename  
  
overall\_df <- rename (overall\_df, c('new\_year' = 'year','Jurisdiction' = 'jurisdiction', 'Average.scale.score' = 'avg\_scale\_score', 'Standard.Error' = 'std\_error'))

It looks like the standard error column is still funky; let's take a look at it

str(overall\_df)

## 'data.frame': 208 obs. of 4 variables:  
## $ year : int 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...  
## $ jurisdiction : Factor w/ 52 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ avg\_scale\_score: int 267 280 283 275 275 286 284 280 263 291 ...  
## $ std\_error : num 1.237 0.962 1.393 1.429 1.325 ...

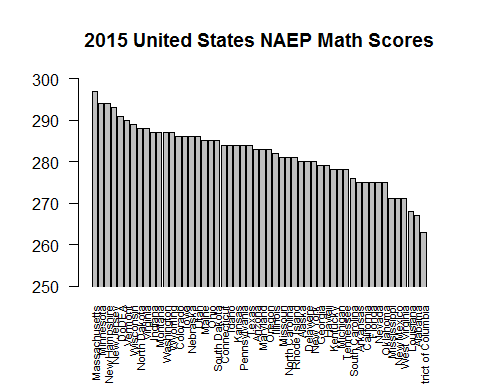
It looks like our data is now good to go

How does Texas rank compared to other states in the country

# order the data by score  
  
analysis\_df <- overall\_df[order(overall\_df$year, overall\_df$avg\_scale\_score, decreasing =TRUE), ]  
  
  
# let'st just look at the 2015 values first:  
  
analysis\_df <- subset(analysis\_df, year == 2015)  
  
# rank  
  
analysis\_df$rank<- NA  
  
analysis\_df$rank <- 1:nrow(analysis\_df)  
  
head(analysis\_df, 53)

## year jurisdiction avg\_scale\_score std\_error rank  
## 23 2015 Massachusetts 297 1.4251418 1  
## 25 2015 Minnesota 294 1.0234131 2  
## 31 2015 New Hampshire 294 0.8990043 3  
## 32 2015 New Jersey 293 1.2362303 4  
## 10 2015 DoDEA 291 0.7241963 5  
## 47 2015 Vermont 290 0.7428193 6  
## 51 2015 Wisconsin 289 1.3076917 7  
## 36 2015 North Dakota 288 0.6657762 8  
## 48 2015 Virginia 288 1.2450256 9  
## 16 2015 Indiana 287 1.1779016 10  
## 28 2015 Montana 287 0.8479167 11  
## 49 2015 Washington 287 1.2823022 12  
## 52 2015 Wyoming 287 0.7221022 13  
## 6 2015 Colorado 286 1.4710870 14  
## 17 2015 Iowa 286 1.1548312 15  
## 29 2015 Nebraska 286 0.7917841 16  
## 46 2015 Utah 286 1.0929459 17  
## 21 2015 Maine 285 0.6917336 18  
## 37 2015 Ohio 285 1.5736991 19  
## 43 2015 South Dakota 285 0.9196745 20  
## 7 2015 Connecticut 284 1.1923837 21  
## 14 2015 Idaho 284 0.8876500 22  
## 18 2015 Kansas 284 1.2846331 23  
## 40 2015 Pennsylvania 284 1.4532845 24  
## 45 2015 Texas 284 1.2119669 25  
## 3 2015 Arizona 283 1.3933827 26  
## 22 2015 Maryland 283 1.2284201 27  
## 39 2015 Oregon 283 1.2024879 28  
## 15 2015 Illinois 282 1.3095851 29  
## 27 2015 Missouri 281 1.2469731 30  
## 35 2015 North Carolina 281 1.5578420 31  
## 41 2015 Rhode Island 281 0.6553583 32  
## 2 2015 Alaska 280 0.9619621 33  
## 8 2015 Delaware 280 0.7456129 34  
## 34 2015 New York 280 1.3600107 35  
## 12 2015 Georgia 279 1.2143485 36  
## 13 2015 Hawaii 279 0.8190231 37  
## 19 2015 Kentucky 278 0.9229194 38  
## 24 2015 Michigan 278 1.3030973 39  
## 44 2015 Tennessee 278 1.7877465 40  
## 42 2015 South Carolina 276 1.2671474 41  
## 4 2015 Arkansas 275 1.4291864 42  
## 5 2015 California 275 1.3250740 43  
## 11 2015 Florida 275 1.3611719 44  
## 30 2015 Nevada 275 0.7334319 45  
## 38 2015 Oklahoma 275 1.3054150 46  
## 26 2015 Mississippi 271 1.1350433 47  
## 33 2015 New Mexico 271 1.0029955 48  
## 50 2015 West Virginia 271 0.8511552 49  
## 20 2015 Louisiana 268 1.3995581 50  
## 1 2015 Alabama 267 1.2370776 51  
## 9 2015 District of Columbia 263 0.9055931 52

# let's look at the ranks of the math scores  
  
barplot(analysis\_df$avg\_scale\_score, names.arg = analysis\_df$jurisdiction, cex.names = 0.7, las = 2, ylim = c(250,300), xpd = FALSE, main = "2015 United States NAEP Math Scores")



It looks like Texas is ranked 25 in 2015. I wonder if that has changed over the past several years?

# need to rank and pull the data together for the previous years:  
  
# create subdfs (not the most efficient, but will be easy to rank)  
  
analysis\_df\_2014 <- subset(overall\_df, year == 2014)  
analysis\_df\_2013 <- subset(overall\_df, year == 2013)  
analysis\_df\_2012 <- subset(overall\_df, year == 2012)  
analysis\_df\_2011 <- subset(overall\_df, year == 2011)  
analysis\_df\_2010 <- subset(overall\_df, year == 2010)  
analysis\_df\_2009 <- subset(overall\_df, year == 2009)  
  
rm(analysis\_df\_2010, analysis\_df\_2012, analysis\_df\_2014)  
  
# now, we need to make sure that the ordering was maintainted  
  
head(analysis\_df\_2013)

## year jurisdiction avg\_scale\_score std\_error  
## 53 2013 Alabama 269 1.339240  
## 54 2013 Alaska 282 0.934323  
## 55 2013 Arizona 280 1.213533  
## 56 2013 Arkansas 278 1.076064  
## 57 2013 California 276 1.178424  
## 58 2013 Colorado 290 1.225326

# looks like it wasn't...let's resort all DFs  
  
analysis\_df\_2013 <- analysis\_df\_2013[order(analysis\_df\_2013$year, analysis\_df\_2013$avg\_scale\_score, decreasing =TRUE), ]  
  
analysis\_df\_2011 <- analysis\_df\_2011[order(analysis\_df\_2011$year, analysis\_df\_2011$avg\_scale\_score, decreasing =TRUE), ]  
  
analysis\_df\_2009 <- analysis\_df\_2009[order(analysis\_df\_2009$year, analysis\_df\_2009$avg\_scale\_score, decreasing =TRUE), ]  
  
# looks like reordering worked well. Now, let's add in the rank column:  
  
analysis\_df\_2013$rank<- NA  
analysis\_df\_2013$rank <- 1:nrow(analysis\_df\_2013)  
  
  
analysis\_df\_2011$rank<- NA  
analysis\_df\_2011$rank <- 1:nrow(analysis\_df\_2011)  
  
analysis\_df\_2009$rank<- NA  
analysis\_df\_2009$rank <- 1:nrow(analysis\_df\_2009)

Now, we should be good to go for a little more analysis. We may want to consolidate all of this information into one dataframe just to be "tidy"

# a very untidy approach, but works...Should really use the merge functionality in the Tidyverse  
  
tidy\_df <- sqldf('Select T1.jurisdiction As jurisdiction,   
T1.rank As "2015\_rank",   
T1.avg\_scale\_score As "2015\_avg\_score",  
T1.std\_error As "2015\_std\_error",  
  
T4.rank As "2013\_rank",   
T4.avg\_scale\_score As "2013\_avg\_score",  
T4.std\_error As "2013\_std\_error",  
  
T3.rank As "2011\_rank",   
T3.avg\_scale\_score As "2011\_avg\_score",  
T3.std\_error As "2011\_std\_error",  
  
T2.rank As "2009\_rank",  
T2.avg\_scale\_score As "2009\_avg\_score",  
T2.std\_error As "2009\_std\_error"  
  
  
From analysis\_df T1  
  
Left Outer Join analysis\_df\_2009 T2   
On T1.jurisdiction = T2.jurisdiction  
  
Left Outer Join analysis\_df\_2011 T3  
On T1.jurisdiction = T3.jurisdiction  
  
Left Outer Join analysis\_df\_2013 T4  
On T1.jurisdiction = T4.jurisdiction')

## Loading required package: tcltk

## Warning: Quoted identifiers should have class SQL, use DBI::SQL() if the  
## caller performs the quoting.

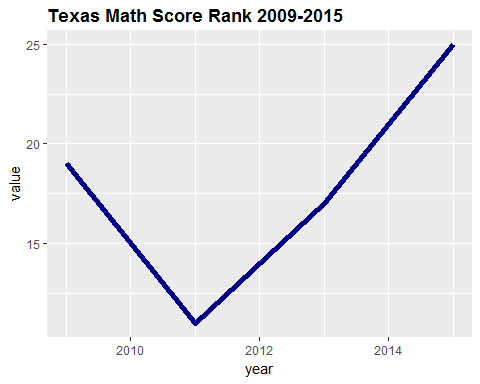
# remove the unused dfs from the environment  
  
rm(analysis\_df\_2009, analysis\_df\_2011, analysis\_df\_2013)

Let's look at the scores for Texas over the past few years

par(mfrow = c(1,2))  
  
texas\_df <- subset(tidy\_df, tidy\_df$jurisdiction == "Texas")  
  
# proly want to melt it   
  
  
texas\_df <- melt(texas\_df)

## Using jurisdiction as id variables

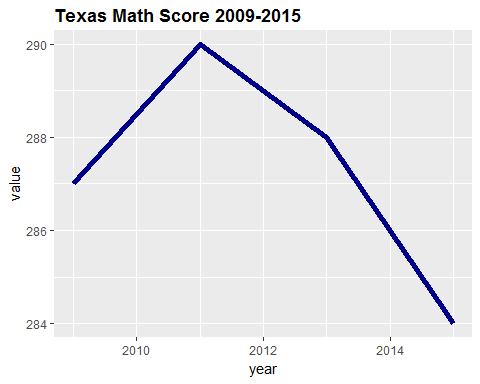
texas\_df <- sqldf('Select \* From texas\_df Where variable Like ("%rank%")')  
  
texas\_df$year <- c(2015, 2013, 2011, 2009)  
  
ggplot(data = texas\_df, aes(x = year, y = value)) + geom\_line(color = 'dark blue', size = 2) + ggtitle('Texas Math Score Rank 2009-2015') + theme(plot.title = element\_text(lineheight = 0.8, face = 'bold'))



# do the same for score  
  
texas\_df <- subset(tidy\_df, tidy\_df$jurisdiction == "Texas")  
  
texas\_df <- melt(texas\_df)

## Using jurisdiction as id variables

texas\_df <- sqldf('Select \* From texas\_df Where variable Like ("%score%")')  
  
texas\_df$year <- c(2015, 2013, 2011, 2009)  
  
ggplot(data = texas\_df, aes(x = year, y = value)) + geom\_line(color = 'dark blue', size = 2) + ggtitle('Texas Math Score 2009-2015') + theme(plot.title = element\_text(lineheight = 0.8, face = 'bold'))



Maybe this is a little misleading, but what it shows is that the math score rankings for Texas have dropped over the past 7 years. So, it looks like the real issue is that the Texas Math score has dropped over the past few years. WE would need to look at some additional information to figure out why this is the case. I wonder if the new score is statistically significantly different from 2009 levels though. Remember that ggplot has automatically truncated the axis

# change since 2009  
  
(subset(tidy\_df, tidy\_df$jurisdiction == "Texas"))$'2015\_avg\_score'

## [1] 284

(subset(tidy\_df, tidy\_df$jurisdiction == "Texas"))$'2009\_avg\_score'

## [1] 287

(subset(tidy\_df, tidy\_df$jurisdiction == "Texas"))$'2015\_avg\_score' /  
(subset(tidy\_df, tidy\_df$jurisdiction == "Texas"))$'2009\_avg\_score' -1

## [1] -0.01045296

# peak score  
  
(subset(tidy\_df, tidy\_df$jurisdiction == "Texas"))$'2011\_avg\_score'

## [1] 290

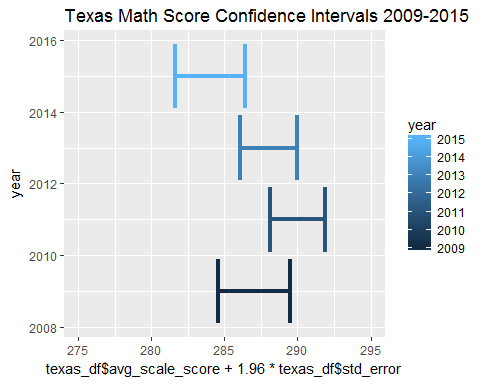
# difference from peak  
  
(subset(tidy\_df, tidy\_df$jurisdiction == "Texas"))$'2015\_avg\_score' /  
(subset(tidy\_df, tidy\_df$jurisdiction == "Texas"))$'2011\_avg\_score' -1

## [1] -0.02068966

The score has only dropped 1% since 2009. Granted, the peak score of 290 has seen a 2% decline. I wonder if the testing procedures changed or something?

Let's see if the difference between the years has a statistically significant difference.

# let's look at error bars for the overlap possibilities usng the overall cleaned frame:  
  
texas\_df = subset(overall\_df, overall\_df$jurisdiction == 'Texas')  
  
# reorder  
  
texas\_df = sqldf('Select \* From texas\_df Order By year')  
  
# now plot the 95% confidence intervals: note that the std. error has already been calculated for us:  
  
  
ggplot(texas\_df, aes(x = year, color = year))+ geom\_errorbar(aes(ymax = texas\_df$avg\_scale\_score + 1.96\*texas\_df$std\_error), ymin = texas\_df$avg\_scale\_score - 1.96\*texas\_df$std\_error, position = 'dodge', lwd = 1.5) + coord\_flip(ylim = c(295, 275)) + ggtitle('Texas Math Score Confidence Intervals 2009-2015')



It looks like the 2011 score is statistically significantly different from the 2015 score. Of course, we are not making any adjustments such as bonferroni for multiple comparisons. It does appear however, that 2015 does show some signs of being significatnly worse for Texas than the other years.