The ***Mendoza Line*** is a term from baseball. Named after Mario Mendoza, it refers to the threshold of incompetent hitting. It is frequently taken to be a batting average of .200, although all the sources I looked at made sure to note that Mendoza's career average was actually a little better: .215.

This post explores a few questions related to the Mendoza line:

- Was Mario Mendoza really so bad as to warrant the expression being named after him?
- How many players fell below the Mendoza line each year?
- What might the Mendoza line look like if we allowed it to change dynamically?

All code for this post can be found here.

(***Caveat:*** I don't know baseball well, so some of the assumptions or conclusions I make below may not be good ones. If I made a mistake, let me know!)

The data

The `Lahman` package on CRAN contains all the baseball statistics from 1871 to 2019. We'll use the `Batting` data frame for statistics and the `People` data frame for player names.

```
library(Lahman)
library(tidyverse)
data(Batting)
data(People)

# add AVG to Batting
Batting$AVG <- with(Batting, H / AB)
```

First, let's look for Mario Mendoza and verify that his batting average is indeed .215:

```
# find Mario Mendoza in People
People %>% filter(nameFirst == "Mario" & nameLast == "Mendoza")
# his ID is mendoma01

Batting %>% filter(playerID == "mendoma01") %>%
  summarize(career_avg = sum(H) / sum(AB))
#   career_avg
# 1  0.2146597
```

Was Mario Mendoza really so bad as to warrant the expression being named after him?

Let's compute the career batting averages for the players and limit our dataset to just the players with at least 1000 at bats in their career:

```
# Batting average for players with >= 1000 AB
avg_df <- Batting %>% group_by(playerID) %>%
  summarize(tot_AB = sum(AB), career_avg = sum(H) / sum(AB)) %>%
  filter(tot_AB >= 1000) %>%
  left_join(People, by = "playerID") %>%
  select(playerID, tot_AB, career_avg, nameFirst, nameLast) %>%
```

```
  arrange(desc(career_avg))
```

Let's look at the top 10 players by batting average: we should see some famous names there! (If not, maybe 1000 ABs is not stringent enough of a criterion to rule out small sample size?)

```
# top 10
head(avg_df, n = 10)
# # A tibble: 10 x 5
#    playerID  tot_AB career_avg nameFirst    nameLast
#
#  1 cobbty01   11436      0.366 Ty           Cobb
#  2 barnero01   2391      0.360 Ross         Barnes
#  3 hornsro01   8173      0.358 Rogers       Hornsby
#  4 jacksjo01   4981      0.356 Shoeless Joe Jackson
#  5 meyerle01   1443      0.356 Levi         Meyerle
#  6 odoulle01   3264      0.349 Lefty        O'Doul
#  7 delahed01   7510      0.346 Ed           Delahanty
#  8 mcveyca01   2513      0.346 Cal          McVey
#  9 speaktr01  10195      0.345 Tris         Speaker
# 10 hamilbi01   6283      0.344 Billy        Hamilton
```

Next, let's look at the bottom 10 players by batting average:

```
# bottom 10
tail(avg_df, n = 10)
# # A tibble: 10 x 5
#    playerID  tot_AB career_avg nameFirst nameLast
#
#  1 seaveto01   1315      0.154 Tom       Seaver
#  2 donahre01   1150      0.152 Red       Donahue
#  3 fellebo01   1282      0.151 Bob       Feller
#  4 grovele01   1369      0.148 Lefty     Grove
#  5 suttodo01   1354      0.144 Don       Sutton
#  6 amesre01    1014      0.141 Red       Ames
#  7 faberre01   1269      0.134 Red       Faber
#  8 perryga01   1076      0.131 Gaylord   Perry
#  9 pappami01   1073      0.123 Milt      Pappas
# 10 frienbo01   1137      0.121 Bob       Friend
```

Those numbers look quite a lot smaller than Mendoza's! Notice also that all of them have ABs just over 1000, my threshold for this dataset. Maybe 1000 ABs is too loose of a condition… But Mendoza only had 1337 ABs, so if we make the condition more stringent (e.g. considering only players with >= 2000 ABs), it's not fair to pick on him…

Among players with >= 1000 ABs, how poor was Mendoza's performance?

```
# How far down was Mario Mendoza?
which(avg_df$playerID == "mendoma01") / nrow(avg_df)
# [1] 0.9630212
```

He's roughly at the 5th quantile of all players for batting average.

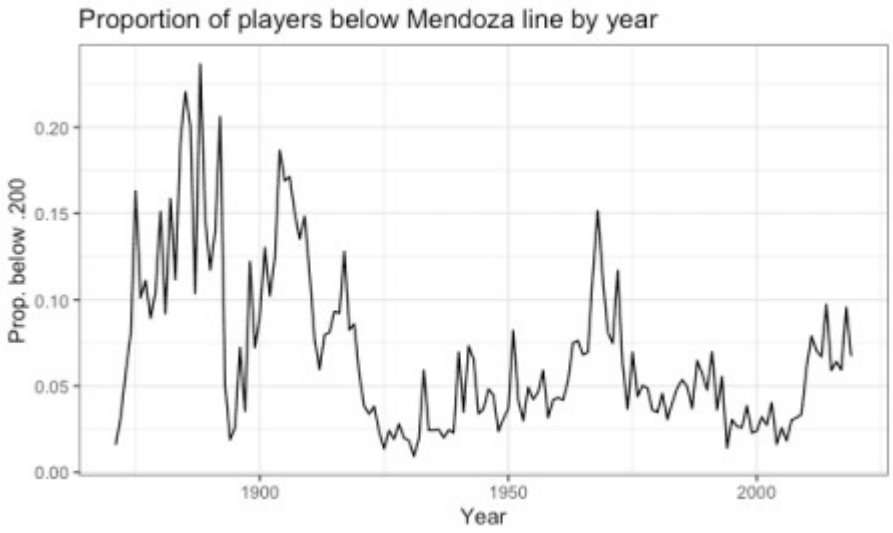
How many players fell below the Mendoza line each year?

For this question, in each season we only consider players who had at least 100 ABs that season.

```
# Look at player-seasons with at least 100 ABs
batting_df <- Batting %>% filter(AB >= 100)
```

The nice thing about the tidyverse is that we can answer the question above with a series of pipes ending in a plot:

```
batting_df %>% group_by(yearID) %>%
  summarize(below200 = mean(AVG < 0.200)) %>%
  ggplot(aes(yearID, below200)) +
  geom_line() +
  labs(title = "Proportion of players below Mendoza line by year",
       x = "Year", y = "Prop. below .200") +
  theme_bw()
```



There is a fair amount of fluctuation, with the proportion of players under the Mendoza line going as high as 24% and as low as 1%. If I had to guess, for the last 50 years or so the proportion seems to fluctuate around 5%.

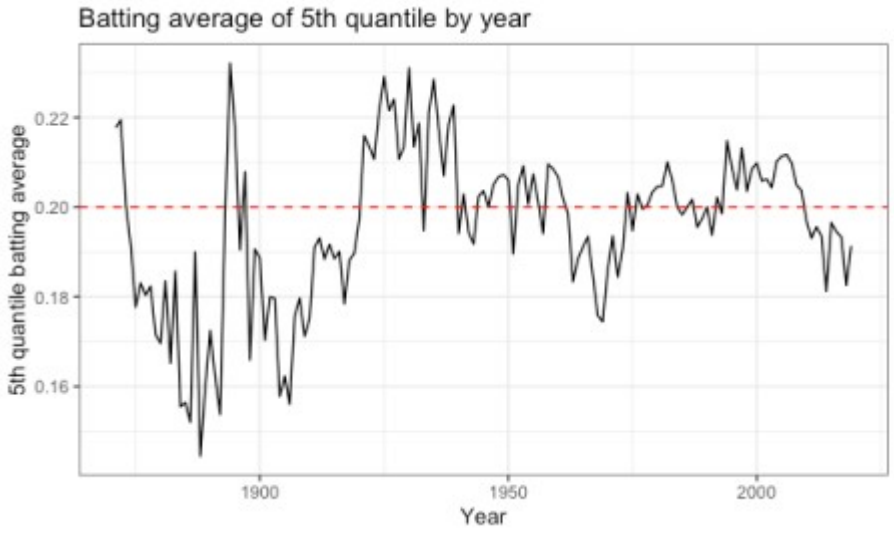
What might the Mendoza line look like if we allowed it to change dynamically?

Instead of defining the Mendoza line as having a batting average below .200, what happens if we define the Mendoza line for a particular season as the batting average of the player at the 5th quantile?

We can answer this easily by summarizing the data using the `quantile` function (dplyr v1.0.0 makes this easy):

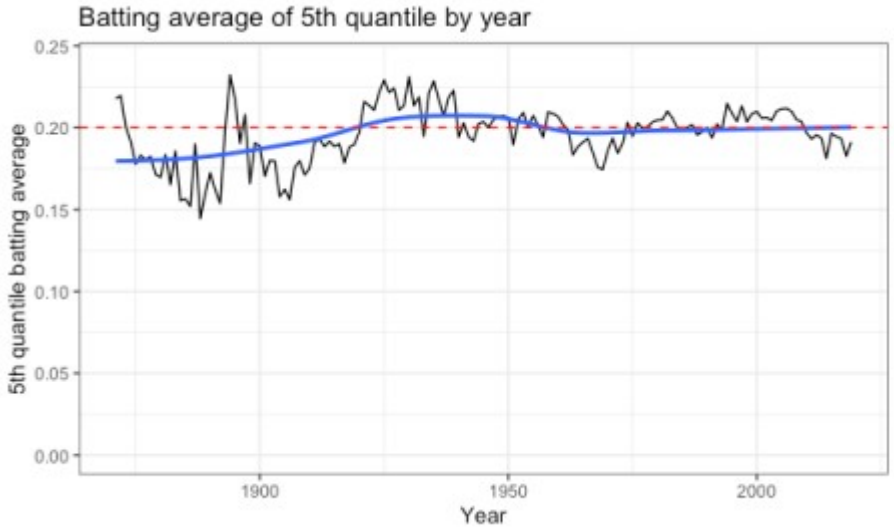
```
batting_df %>% group_by(yearID) %>%
  summarize(bottom5 = quantile(AVG, 0.05)) %>%
  ggplot(aes(yearID, bottom5)) +
  geom_line() +
```

```
  geom_hline(yintercept = c(0.2), color = "red", linetype = "dashed") +
  labs(title = "Batting average of 5th quantile by year",
       x = "Year", y = "5th quantile batting average") +
  theme_bw()
```



That looks like a lot of fluctuation, but if you look closely at the y-axis, you'll see that the values hover between 0.14 and 0.24. Here is the same line graph but with zero included on the y-axis and a loess smoothing curve:

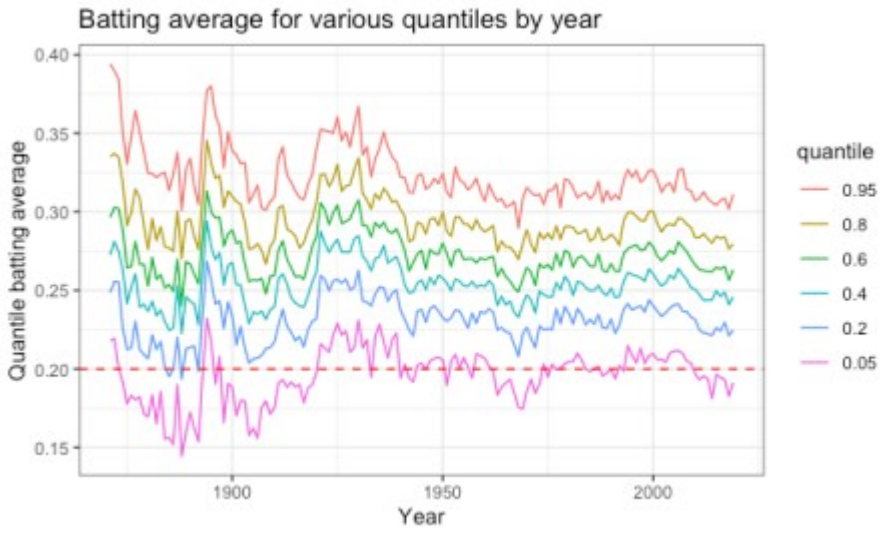
```
batting_df %>% group_by(yearID) %>%
  summarize(bottom5 = quantile(AVG, 0.05)) %>%
  ggplot(aes(yearID, bottom5)) +
  geom_line() +
  geom_smooth(se = FALSE) +
  geom_hline(yintercept = c(0.2), color = "red", linetype = "dashed") +
  scale_y_continuous(limits = c(0, 0.24)) +
  labs(title = "Batting average of 5th quantile by year",
       x = "Year", y = "5th quantile batting average") +
  theme_bw()
```



I'm not sure how much to trust that smoother, but it comes awfully close to the Mendoza line!

For completeness, here are the lines representing the batting averages for players at various quantiles over time:

```
batting_df %>% group_by(yearID) %>%
  summarize(AVG = quantile(AVG, c(0.05, 1:4 / 5, 0.95)),
            quantile = c(0.05, 1:4 / 5, 0.95)) %>%
  mutate(quantile = factor(quantile, levels = c(0.95, 4:1 / 5, 0.05)))
%>%
  ggplot(aes(x = yearID, y = AVG, col = quantile)) +
  geom_line() +
  geom_hline(yintercept = c(0.2), color = "red", linetype = "dashed") +
  labs(title = "Batting average for various quantiles by year",
       x = "Year", y = "Quantile batting average") +
  theme_bw()
```



At a glance the higher quantile lines look just like vertical translations of the 5th quantile line, suggesting that across years the entire distribution of batting averages shifts up or down (not just parts of the distribution).
