Lesson_35_Boardsheet

Kevin Cummiskey

4/26/2020

DiMaggio's Streak

In 1941, Joe DiMaggio of the New York Yankees hit safely (got at least one hit) in 56 straight games. The streak began on May 15th, 1941 when DiMaggio went 1-for-4 against the Chicago White Sox and ended on July 17th when he went 0-for-3 with a walk against the Cleveland Indians.

The following example comes from Chapter 10 of *Analyzing Baseball Data with R* by Marchi, Albert, and Baumer.

```
library(tidyverse)
library(knitr)

# game-by-game records for DiMaggio's 1941 season
file = "dimaggio.1941.csv"
path = "https://raw.githubusercontent.com/maxtoki/baseball_R/master/data/"
joe <- read_csv(file = paste(path,file,sep=""))

joe %>%
  select(Date, Opp, PA, H) %>%
  head(10)
```

```
## # A tibble: 10 x 4
##    Date   Opp      PA     H
##    <chr>  <chr> <dbl> <dbl>
##  1 14-Apr WSH       4     2
##  2 15-Apr PHA       4     2
##  3 16-Apr PHA       5     4
##  4 17-Apr PHA       5     2
##  5 18-Apr WSH       4     1
##  6 19-Apr WSH       5     1
##  7 20-Apr PHA       6     3
##  8 21-Apr PHA       6     4
```

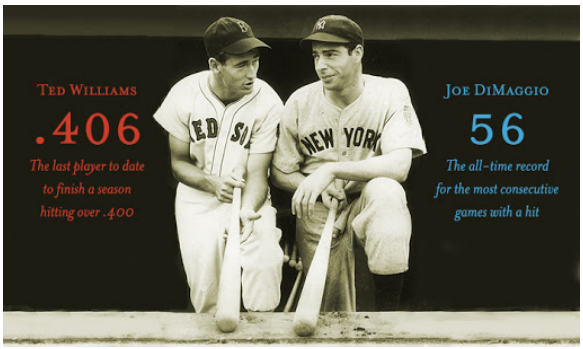


Figure 1: Ted Williams and Joe DiMaggio

```
##  9 22-Apr PHA       4     0
## 10 23-Apr BOS       5     0
```

```
#add a variable indicating whether or not he got a hit in the game
joe %>%
  mutate(HIT = ifelse(H > 0, 1,0)) -> joe

joe %>%
  count(HIT) %>%
  kable(caption = "Number of Games by whether DiMaggio got at least one hit (1) or not (0).")
```

Table 1: Number of Games by whether DiMaggio got at least one hit (1) or not (0).

HIT	n
0	25
1	114

```
#function to calculate streak
streaks <- function(y){
  x <- rle(y)
  class(x) <- "list"
  return(as_tibble(x))
}

joe %>%
  pull(HIT) %>%
  streaks() %>%
  filter(values == 1) %>%
  pull(lengths)
```

```
##  [1]  8  3  2  1  3 56 16  4  2  4  7  1  5  2
```

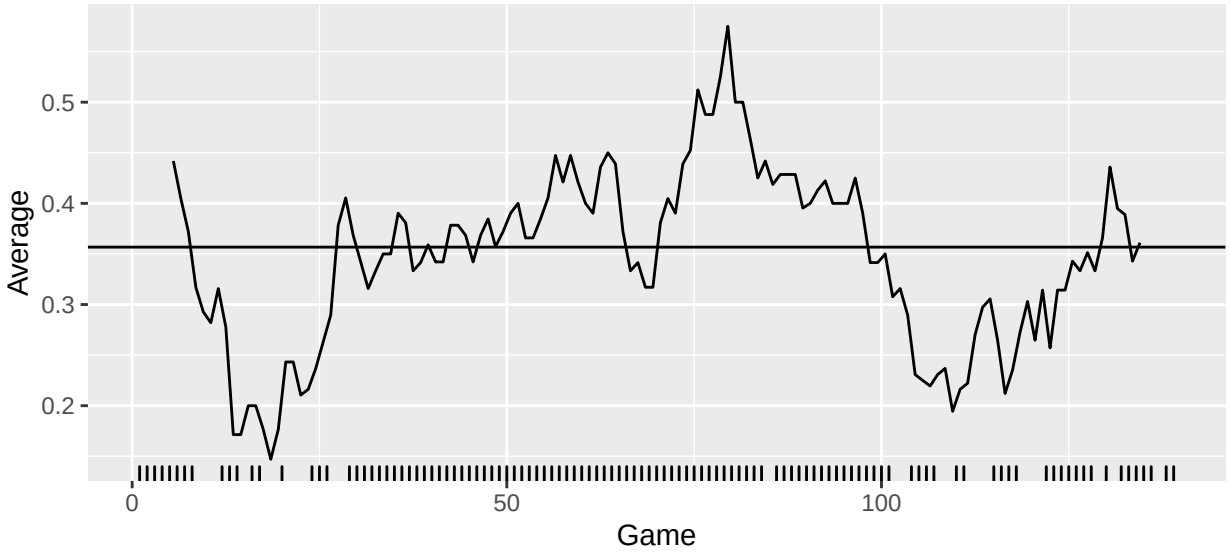
Moving Average

```
library(zoo)
moving_average <- function(df, width){
  N <- nrow(df)
  df %>%
    transmute(Game = rollmean(1:N, k = width, fill = NA),
              Average = rollsum(H, width, fill = NA)/
                rollsum(AB, width, fill = NA))
}

joe_ma <- moving_average(joe,10)

joe_ma %>%
  ggplot(aes(x = Game, y = Average)) +
  geom_line() +
```

```
  geom_hline(data = summarize(joe, bavg = sum(H)/sum(AB)),
             aes(yintercept = bavg)) +
  geom_rug(data = filter(joe, HIT == 1),
           aes(Rk, .3*HIT), sides = "b")
```



How unusual was DiMaggio's streak for a player who hit safely in 114 of 139 games?

- Let's review Father Costa's approach.

https://www.mlb.com/news/baseball-hitting-streaks-c265612772

- Here's another approach using simulation.

For a player who hit safely in 114 of 139 games, let's see how unusual it would be for him to have a streak of 56 or more consecutive games using simulation. The follow code shuffles the *HIT* column and returns the longest streak for each of r simulated seasons.

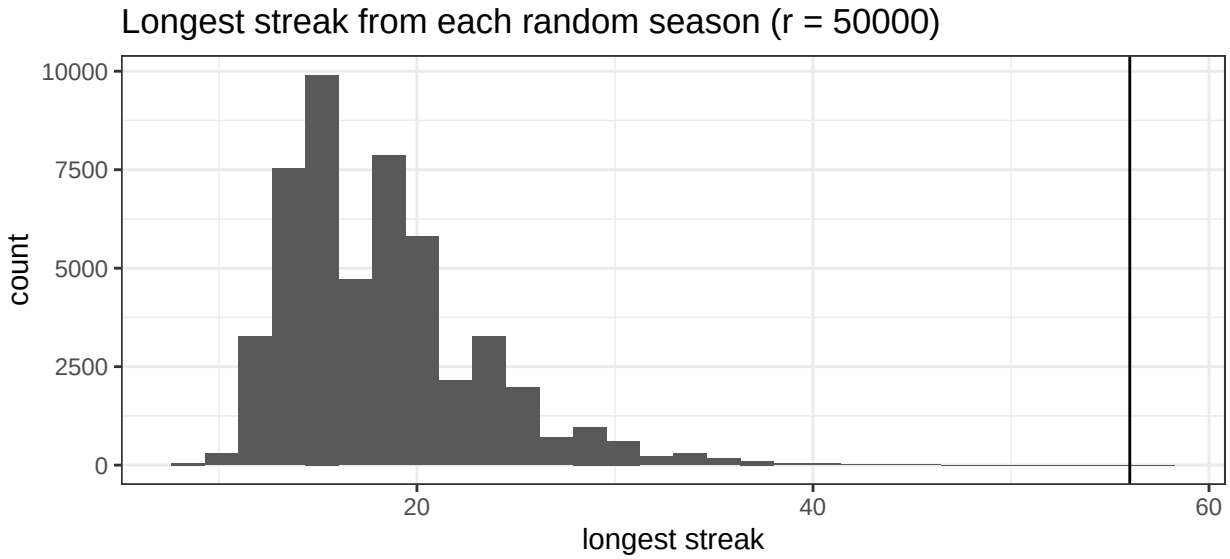
```
# this function shuffles the y vector and returns the longest streak
random_mix <- function(y){
  y %>%
    sample() %>%
    streaks() %>%
    filter(values == 1) %>%
    arrange(-lengths) %>%
    head(1) %>%
    pull(lengths)
}

r = 50000 #number of replications

#run similulation experiment
```

```r
#you can think of replication as a for loop
joe_random <- replicate(r, random_mix(joe$HIT))

sim.result <- tibble(streak.long = joe_random)
sim.result %>%
  ggplot(aes(x = streak.long)) +
  geom_histogram() +
  theme_bw() +
  labs(title = paste("Longest streak from each random season (r = ",r,")", sep = ""),
       x = "longest streak") +
  geom_vline(xintercept = 56)
```




```r
sum(joe_random >= 56)/r
```

```
## [1] 2e-05
```

How do we interpret the graph above?

Should we interpret the probability above as the probability someone breaks the record in the next 50 years?

Streakiness statistic