[r-bloggers.com](https://www.r-bloggers.com/riddler-can-you-beat-mlb-recods/)

**Riddler: Can You Beat MLB Recods?**

8-10 minutes

[This article was first published on [**Posts | Joshua Cook**](https://joshuacook.netlify.app/post/riddler-beat-mlb-records/), and kindly contributed to [R-bloggers](https://www.r-bloggers.com/)]. (You can report issue about the content on this page [here](https://www.r-bloggers.com/contact-us/))

Want to share your content on R-bloggers? [click here](https://www.r-bloggers.com/add-your-blog/) if you have a blog, or [here](https://r-posts.com/) if you don't.

**FiveThirtyEight’s Riddler Express**

[link](https://fivethirtyeight.com/features/can-the-hare-beat-the-tortoise/)

From Taylor Firman comes an opportunity to make baseball history:

This year, Major League Baseball announced it will play a shortened  
60-game season, as opposed to the typical 162-game season. Baseball is  
a sport of numbers and statistics, and so Taylor wondered about the  
impact of the season’s length on some famous baseball records.

Some statistics are more achievable than others in a shortened season.  
**Suppose your true batting average is .350, meaning you have a 35  
percent chance of getting a hit with every at-bat. If you have four  
at-bats per game, what are your chances of batting at least .400 over  
the course of the 60-game season?2 And how does this compare to your  
chances of batting at least .400 over the course of a 162-game season?**

**Plan**

This riddle should be pretty straight forward to solve statistically and  
with simulations, so I will do both.

**Setup**

knitr::opts\_chunk$set(echo = TRUE, comment = "#>", cache = FALSE, dpi = 400)

library(mustashe)

library(tidyverse)

library(conflicted)

# Handle any namespace conflicts.

conflict\_prefer("filter", "dplyr")

conflict\_prefer("select", "dplyr")

# Default 'ggplot2' theme.

theme\_set(theme\_minimal())

# For reproducibility.

set.seed(123)

**Statistical solution**

This is a simple binomial system: at each at bat, the player either gets  
a hit or not. If their real batting average is 0.350, that means the  
probability of them getting a hit during each at bat is 35%. Thus,  
according the the Central Limit Theorem, given a sufficiently large  
number of at bats, the frequency of a hit should be 35%. This is because  
the distribution converges towards the mean. However, at smaller  
sample-sizes, the distribution is more broad, meaning that the observed  
batting average has a greater chance of being further away from the true  
value.

First, let’s answer the riddle. The solution is just the probability of  
observing a batting average of 0.400 or greater. The first value is  
computed using dbinom() and the second cumulative probability is  
calculated using pbinom(), setting lower.tail = FALSE to get the  
tail above 0.400.

num\_at\_bats <- 60 \* 4

real\_batting\_average <- 0.350

target\_batting\_average <- 0.400

prob\_at\_400 <- dbinom(x = target\_batting\_average \* num\_at\_bats,

size = num\_at\_bats,

prob = real\_batting\_average)

prob\_above\_400 <- pbinom(q = target\_batting\_average \* num\_at\_bats,

size = num\_at\_bats,

prob = real\_batting\_average,

lower.tail = FALSE)

prob\_at\_400 + prob\_above\_400

#> [1] 0.06083863

**Under the described assumptions, there is a 6.1% chance of reaching a  
batting average of 0.400 in the shorter season.**

For comparison, the chance for a normal 162-game season is calculated  
below. Because $0.400 \times 162 \times 4$ is a non-integer value, an  
exact 0.400 batting average cannot be obtained. Therefore, only the  
probability of a batting average greater than 0.400 needs to be  
calculated.

num\_at\_bats <- 162 \* 4

real\_batting\_average <- 0.350

target\_batting\_average <- 0.400

prob\_above\_400 <- pbinom(q = target\_batting\_average \* num\_at\_bats,

size = num\_at\_bats,

prob = real\_batting\_average,

lower.tail = FALSE)

prob\_above\_400

#> [1] 0.003789922

Over 162 games, there is a 0.4% chance of achieving a batting average of  
at least 0.400.

**Simulation**

The solution to this riddle could also be found by simulating a whole  
bunch of seasons with the real batting average of 0.350 and then just  
counting how frequently the simulations resulted in an observed batting  
average of 0.400.

A single season can be simulated using the rbinom() function where n  
is the number of seasons to simulate, size takes the number of at  
bats, and prob takes the true batting average. The returned value is a  
sampled number of hits (“successes”) over the season from the binomial  
distribution.

The first example shows the observed batting average from a single  
season.

num\_at\_bats <- 60 \* 4

real\_batting\_average <- 0.350

target\_batting\_average <- 0.400

rbinom(n = 1, size = num\_at\_bats, prob = real\_batting\_average) / num\_at\_bats

#> [1] 0.3375

The n = 1 can just be replaced with a large number to simulate a bunch  
of seasons. The average batting average over these seasons should be  
close to the true batting average.

n\_seasons <- 1e6 # 1 million simulations.

sim\_res <- rbinom(n = n\_seasons,

size = num\_at\_bats,

prob = real\_batting\_average)

sim\_res <- sim\_res / num\_at\_bats

# The average batting average is near the true batting average of 0.350.

mean(sim\_res)

#> [1] 0.3500121

The full distribution of batting averages over the 1 million simulations  
is shown below.

tibble(sims = sim\_res) %>%

ggplot(aes(x = sims)) +

geom\_density(color = "black", fill = "black", adjust = 2,

alpha = 0.2, size = 1.2, ) +

geom\_vline(xintercept = target\_batting\_average,

color = "tomato", lty = 2, size = 1.2) +

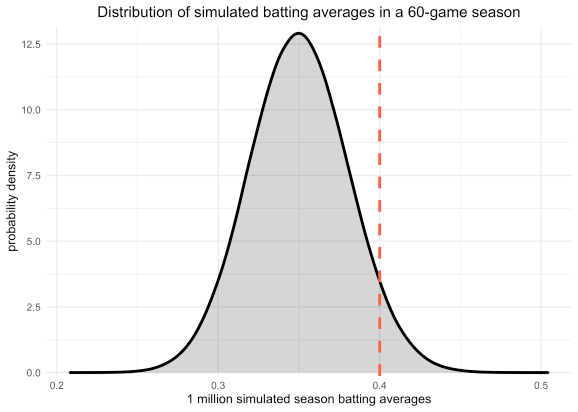
scale\_y\_continuous(expand = expansion(mult = c(0.01, 0.02))) +

theme(plot.title = element\_text(hjust = 0.5)) +

labs(x = "1 million simulated season batting averages",

y = "probability density",

title = "Distribution of simulated batting averages in a 60-game season")



The answer from the simulation is pretty close to the actual answer.

mean(sim\_res >= 0.40)

#> [1] 0.060813

One last visualization I want to do demonstrates why the length of the  
season matters to the distribution. Instead of using rbinom() to  
simulate the number of successes over the entire season, I use it below  
to simulate a season’s-worth of individual at bats, returning a vector  
of 0’s and 1’s. I then plotted the cumulative number of hits at each at  
bat and colored the line by the running batting average.

The coloring shows how the batting average was more volatile when there  
were fewer at bats.

sampled\_at\_bats <- rbinom(60\*4, 1, 0.35)

tibble(at\_bat = sampled\_at\_bats) %>%

mutate(i = row\_number(),

cum\_total = cumsum(at\_bat),

running\_avg = cum\_total / i) %>%

ggplot(aes(x = i, y = cum\_total)) +

geom\_line(aes(color = running\_avg), size = 1.2) +

scale\_color\_viridis\_c() +

theme(plot.title = element\_text(hjust = 0.5),

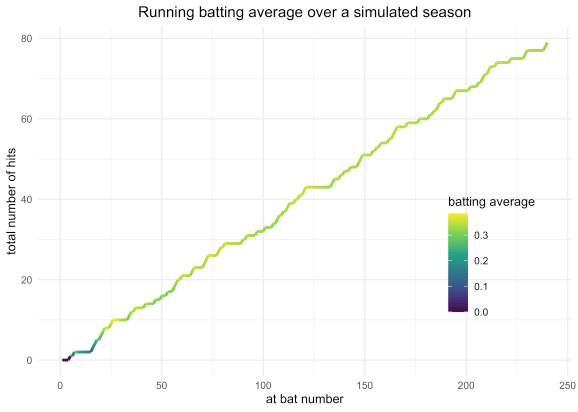
legend.position = c(0.85, 0.35)) +

labs(x = "at bat number",

y = "total number of hits",

color = "batting average",

title = "Running batting average over a simulated season")



The following two plots do the same analysis many times to simulate many  
seasons and color the lines by whether or not the final batting average  
was at or above 0.400. As there are more games, the running batting  
averages, which are essentially biased random walks, regress towards the  
true batting average. (Note that I had to do 500 simulations for the  
162-game season to get any simulations with a final batting average  
above 0.400.)

simulate\_season\_at\_bats <- function(num\_at\_bats) {

sampled\_at\_bats <- rbinom(num\_at\_bats, size = 1, prob = 0.35)

tibble(result = sampled\_at\_bats) %>%

mutate(at\_bat = row\_number(),

cum\_total = cumsum(result),

running\_avg = cum\_total / at\_bat)

}

tibble(season = 1:100) %>%

mutate(season\_results = map(season, ~ simulate\_season\_at\_bats(60 \* 4))) %>%

unnest(season\_results) %>%

group\_by(season) %>%

mutate(

above\_400 = 0.4 <= running\_avg[which.max(at\_bat)],

above\_400 = ifelse(above\_400, "BA ≥ 0.400", "BA < 0.400")

) %>%

ungroup() %>%

ggplot(aes(x = at\_bat, y = cum\_total)) +

geom\_line(aes(group = season, color = above\_400,

alpha = above\_400),

size = 0.8) +

geom\_hline(yintercept = 0.4 \* 60 \* 4,

color = "tomato", lty = 2, size = 1) +

scale\_color\_manual(values = c("grey50", "dodgerblue")) +

scale\_alpha\_manual(values = c(0.1, 1.0), guide = FALSE) +

scale\_x\_continuous(expand = c(0, 0)) +

theme(plot.title = element\_text(hjust = 0.5),

plot.subtitle = element\_text(hjust = 0.5),

legend.position = c(0.85, 0.25)) +

labs(x = "at bat number",

y = "total number of hits",

color = NULL,

title = "Running batting averages over simulated 60-game seasons",

subtitle = "Blue lines indicate a simulation with a final batting average of at least 0.400.")

