Assignment 3: James-Stein Estimator for Batting Averages

Due Friday, October 17 at 11:59 PM

In this homework assignment you will replicate the results (using 2024 MLB data) of the 1977 paper "Stein's Paradox in Statistics which shows that the James-Stein estimator of batting average for players after 45 at-bats is a better predictor of their end of season batting average than the observed average.

To complete this assignment you will use the datasets called `players.csv` and `at-bats.csv` in the Canvas assignments folder. This data was extracted from the official MLB API and graciously provided to us for use by Austin Harcarik.

Submit your assignment to Gradescope as a PDF file generated from your R Markdown. Make sure your PDF includes all code, outputs, and written answers to questions. All code and outputs used to answer the written questions must be included in the PDF for full credit.

Load the data [2 points]

1 (2 points). Load in `players.csv` and `at_bats.csv`. How many rows and columns are in each dataset? What is the unit of observation for each dataset?

```
players <- read.csv("/Users/atharvvani/Documents/SDS375/players.csv")
at_bats <- read.csv("/Users/atharvvani/Documents/SDS375/at_bats.csv")
```

In the players dataset, there are 17843 observations and 18 columns. The unit of observation for the players dataset, is an entry of a player and the columns represent information about that specific player. In the at_bats dataset, there are 182903 observations and 31 columns. The unit of observation for the at_bats dataset is the event/play where a player is at bat and the columns represent information about that play and what happened.

Calculate batting average [14 points]

2 (2 points). We will create a column that we will be able to aggregate to calculate batting average. First, we need to filter out events that do not count as an at-bat. Filter out the following values for `event_type`:

"walk", "caught_stealing_2b", "sac_bunt", "intent_walk", "sac_fly", "pickoff_1b", "caught_stealing_3b", "pickoff_caught_stealing_2b", "catcher_interf", "stolen_base_2b", "pickoff_2b", "pickoff_caught_stealing_home", "pickoff_caught_stealing_3b", "caught_stealing_home", "pickoff_3b", "sac_fly_double_play", "wild_pitch", "hit_by_pitch"

Now to the filtered at_bats dataset, add a new column called `ba` that equals:

- 0 if event_type is one of the following: "field_out", "strikeout", "force_out", "strikeout_double_play", "double_play", "triple_play", "other_out", "fielders_choice", "fielders_choice_out", "field_error", "grounded_into_double_play"
- 1 if event_type is one of the following: "single", "double", "home_run", "triple"

```
exclude_events <- c(
  "walk", "caught_stealing_2b", "sac_bunt", "intent_walk", "sac_fly",
  "pickoff_1b", "caught_stealing_3b", "pickoff_caught_stealing_2b",
  "catcher_interf", "stolen_base_2b", "pickoff_2b",
  "pickoff_caught_stealing_home", "pickoff_caught_stealing_3b",
  "caught_stealing_home", "pickoff_3b", "sac_fly_double_play",
```

```
  "wild_pitch", "hit_by_pitch"
)

at_bats_filtered <- subset(at_bats, !(event_type %in% exclude_events))

at_bats_filtered$ba <- ifelse(
  at_bats_filtered$event_type %in% c("single", "double", "home_run", "triple"), 1,
  ifelse(at_bats_filtered$event_type %in% c(
    "field_out", "strikeout", "force_out", "strikeout_double_play", "double_play",
    "triple_play", "other_out", "fielders_choice", "fielders_choice_out",
    "field_error", "grounded_into_double_play"
  ), 0, NA)
)
```

3 (6 points). Create a new table called `end_season_ba` with each batter's end of season batting average (i.e., the average of your new `ba` column) and the number of at-bats in the season. Filter out players with fewer than 400 at-bats. Then join this new table to the data you read in from `players.csv` to get each batter's position. Filter `end_season_ba` to include only batters whose primary position is shortstop ("SS").

```
end_season_ba <- at_bats_filtered %>%
  group_by(batter_id) %>%
  summarise(
    at_bats = n(),
    batting_avg = mean(ba, na.rm = TRUE)
  ) %>%
  filter(at_bats >= 400)

# Join with players dataset to get positions
end_season_ba <- end_season_ba %>%
  left_join(players, by = c("batter_id" = "id"))

# Filter to include only shortstops (SS)
end_season_ba <- end_season_ba %>%
  filter(primary_position_abbreviation == "SS")
```

4 (6 points). Create a new table called `ba45_df` with each of the batter's batting average after their first 45 at-bats (hint: use `slice()` to get the first 45 at-bats for each player and make sure you order the dataset by time to get the *first* 45). Only include players that were included in `end_season_ba` (i.e., only include shortstops that have 400 or more at-bats by the end of the season).

```
ba45_df <- at_bats_filtered %>%
  arrange(batter_id, start_time) %>%          # ensure time order
  group_by(batter_id) %>%
  slice_head(n = 45) %>%                      # take first 45 at-bats
  summarise(
    ba45 = mean(ba, na.rm = TRUE),            # average over first 45
    at_bats = n()
  ) %>%
  inner_join(end_season_ba %>% select(batter_id), by = "batter_id") %>%
  ungroup()
```

Calculate the James-Stein estimator [6 points]

5 (6 points). Now let's calculate the James-Stein estimator for batting average and add it as a column to the previous dataframe. First, calculate the grand average of player batting averages after 45 at-bats, the

number of shortstops (with more than 400 at-bats by the end of the season), and the shrinkage factor, c. What value do you get for c? Create a new column in `ba45_df` that gives the James-Stein estimator for each shortstop's batting average.

```
grand_mean <- mean(ba45_df$ba45, na.rm = TRUE)
m <- nrow(ba45_df)
s2 <- var(ba45_df$ba45, na.rm = TRUE)
sigma2 <- grand_mean * (1 - grand_mean) / 45

# James-Stein shrinkage factor + ensure non-negative
c <- 1 - ((m - 3) * sigma2 / sum((ba45_df$ba45 - grand_mean)^2))
c <- max(0, c)

# Add JS estimate
ba45_df <- ba45_df %>%
  mutate(js_estimate = grand_mean + c * (ba45 - grand_mean))

cat("Shrinkage factor (c):", round(c, 4), "\n")
```

```
## Shrinkage factor (c): 0.3902
```

Evaluation [8 points]

6 (4 points). Join the dataframes `ba45_df` and `end_season_ba` by batter id. Use this joined dataframe to calculate the mean-squared error (MSE) between the observed batting average after the first 45 at-bats and the end of season batting average. Calculate the same for the James-Stein estimator. Which has a lower MSE?

```
joined_df <- ba45_df %>%
  left_join(end_season_ba, by = "batter_id")

# Calculate MSEs
mse_ba45 <- mean((joined_df$ba45 - joined_df$batting_avg)^2, na.rm = TRUE)
mse_js   <- mean((joined_df$js_estimate - joined_df$batting_avg)^2, na.rm = TRUE)

# Display results
cat("MSE (Raw 45-AB average):", round(mse_ba45, 6))
```

```
## MSE (Raw 45-AB average): 0.006417
```

```
cat("MSE (James-Stein estimate):", round(mse_js, 6))
```

```
## MSE (James-Stein estimate): 0.001491
```

The calculated MSE for the first 45 at-bats and the end of season batting average is 0.0064. The calculated MSE for the JS estimator is 0.0015. Comparing these values, the lower MSE is the MSE for the James-Stein estimator.

7 (4 points). Create a plot of the observed batting average after 45 at-bats versus the end of season batting average. Overlay a plot, in a different color, of the James-Stein estimate of batting average versus the end of season batting average.

```
joined_df <- ba45_df %>%
  left_join(end_season_ba %>% select(batter_id, batting_avg), by = "batter_id")

ggplot(joined_df, aes(x = batting_avg, y = ba45)) +
  geom_point(color = "steelblue", alpha = 0.6, size = 2) +
  geom_point(aes(y = js_estimate), color = "darkorange", alpha = 0.6, size = 2) +
```

```
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "gray40") +
  labs(
    title = "Observed vs JS Estimates of Batting Average",
    x = "End-of-Season Batting Average",
    y = "Batting Average After 45 At-Bats"
  )
```