Stat 340 Group Progress Report

11/11/2021

Matthew Chiang: 907 723 8120 (Mchiang7)

Ishaan Backliwal: 908 134 7719 (backliwal)

Jordan Livingston: 908 132 1151 (jlivingston4)

Eric Dietze: 907 935 8843 (edietze)

Vu Pham: 907 808 5595 (vmpham2)

Description of Data

Baseball is an American sport played between two teams. One team plays defense and has a player (pitcher) try to throw the ball past the other team's batter who tries to hit the ball. If the batting team successfully hits the ball enough times they can score runs (points). We will use data sets from baseball-reference.com which contain team and individual statistics. This database also offers data 'splits', which show comparisons between home and away games. For our primary analysis, we will be focusing on the Houston Astros 2017 Team Batting Splits and the 2017 Individual Player Batting Splits. These data sets outline various different statistics of the Houston Astros 2017 season in which it was confirmed that they had cheated in home games. We will use the 'splits' data sets to determine a difference of statistics between home and away games. This database also contains data on the Astros for different seasons and different teams. Depending on the results of our initial analysis, we will compare the players on the roster of the 2017 Astros to previous seasons to determine a difference in performance before and after cheating. Our statistical question is whether or not the 2017 Houston Astros cheated. Two years after the alleged cheating occured, reports from a player who was on the 2017 Astros alleged that there were cameras positioned in the Astros home field stadium, Minute Maid Park in Houston, that could see the relayed signs between the pitcher and catcher. Note: the catcher communicates to the pitcher what pitch the pitcher should throw based on a series of hand movements in between pitches. The Houston Astros camera could see these hand signals and people were in charge of watching the film and decoding the symbols to determine what pitch was coming. Once a pitch was shown, members of the Astros would relay what pitch was coming to the batter through their own series of symbols (most infamously a banging of a garbage can).

The Astros only had this technology at their Home Stadium, so they should have higher performance at home compared to on the road. Note: this sign decoding should reason that only their batting improved at home and not their pitching. We used several stats for measuring their performance.

Statistical Questions

- Is there a noticeable difference between the batting splits for home and away games of the Houston Astros team during the 2017 season?
- How does this difference, if any, compare to other teams' performance in the MLB during the 2017 season?

- Is there sufficient statistical evidence to suggest the Houston Astros benefited from cheating in the 2017 season?
  - If there is, how much may the cheating have affected their season? And what would their season have looked like if they didn't cheat?

Why We Chose This Dataset

- This has been a controversy within the baseball community regarded to be one of the biggest cheating scandals in the sport's history.
- In 2017 the Houston Astros won the world series. However, it came out in November of 2019 that they were using technology to steal signs during home games.
- To us, this data set is interesting because it contains (in detail) records of each game, each player, and almost all variables that could take place in this event. So, this dataset is very in depth and useful for testing hypotheses thoroughly regarding the incident. Like they said "Numbers don't lie"!

Variables

Below is a list of some important variables in our dataset:

Name	Abbr.	Description
On base and slugging percentage	OPS	Measures a players On base Percentage (percentage of At bats a player has gotten on base) and a players slugging percentage (a weighted batting average)
On base and slugging percentage (Player)	tOPS+	This is adjusted so that 100 is the team average, so if tOPS is less than 100, the batters did worse, and if it is higher than 100, the batters did better.
On base and slugging percentage (Team and Player)	sOPS+	This is adjusted so that 100 is the league average, so if sOPS is less than 100, the batters did worse than the league average and if it is higher than 100, the batters did better.
Runs Batted In	RBI	The number of runs a player has created by hitting a player home. Weights a players higher
Hits	H	When a batter reaches base without doing so via error or fielder's choice.
Homeruns	HR	Scored when the ball is hit in such a way that the batter is able to circle the bases and reach home safely.
Walks	W	Occurs when a pitcher throws four pitches outside of the strike zone, none of which are swung at by the batter.
Batting Average	BA	Percentage of At Bats a player gets a hit.

Loading Data

```r
# will store the data for all teams in the 2017 season
season_2017 <- NA
data_empty <- TRUE

# Getting initial path data
data_path <- "./data"
data_dirs <- list.files(data_path)
data_dirs <- data_dirs[!str_detect(data_dirs, ".csv")]

# iteratively get all csv file paths and store them
for(path in data_dirs){
```

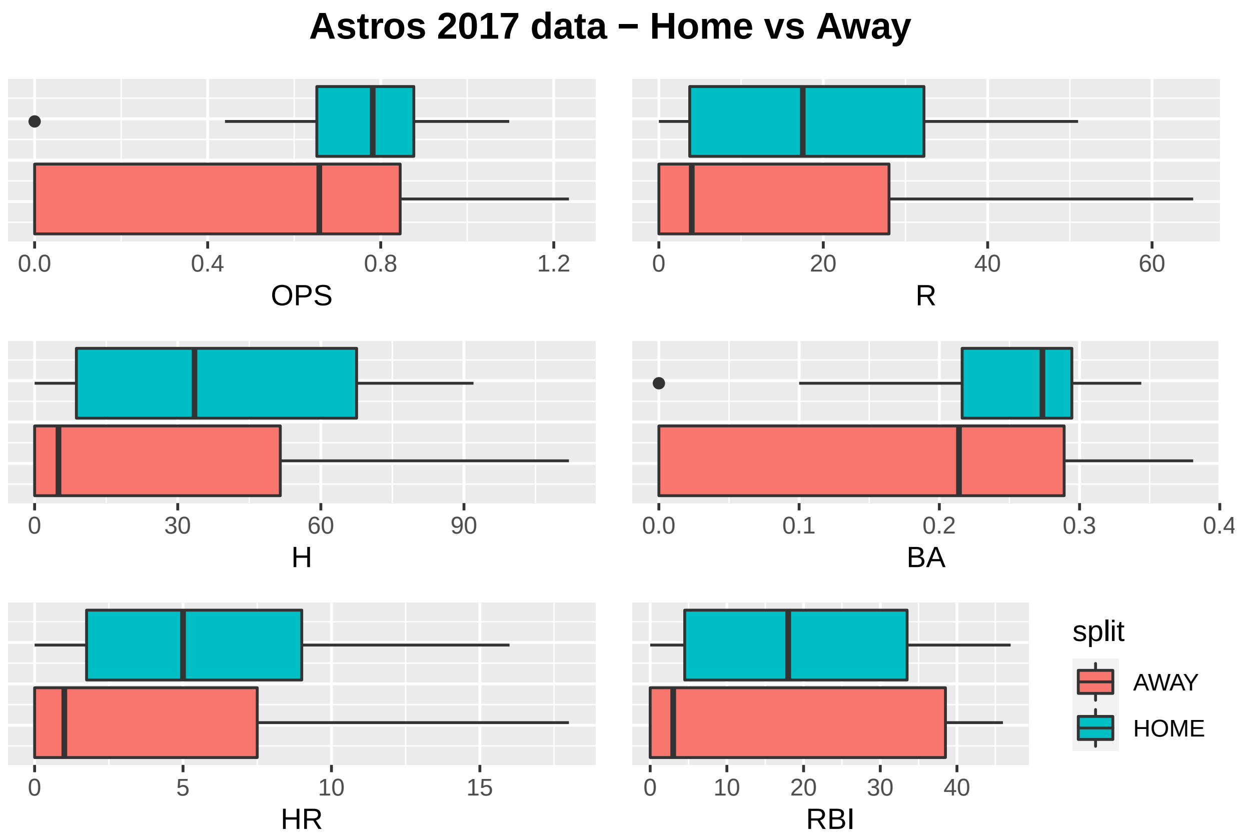
```r
  # gets the directory for a spefic teams csv files
  team_dir <- paste(data_path, path, sep="/")
  csv_list <- list.files(team_dir)

  # steps through csv files, loads and edits them as dataframes
  for(csv in csv_list){
    # get full path, and other information from csv name
    full_path <- paste(team_dir, csv, sep="/")
    split <- ifelse(str_detect(csv, "home"), "HOME", "AWAY")
    team_name <- str_split_fixed(team_dir, pattern="data/", 2)[2]

    # Read in data
    df <- read.csv(full_path)%>%
      mutate(Name = str_split_fixed(Name, "\\\\", 2)[,1]) %>%
      mutate(split = split, team = team_name)
    df <- df[ -length(df[,1]), ]

    # binds all teams data together
    if(data_empty){
      season_2017 <- df
      data_empty <- FALSE
    }else {
      season_2017 <- rbind(season_2017, df, make.row.names=TRUE)
    }
  }
}
astros_2017 <- season_2017 %>%
  filter(team=="astros", AB != 0)
```

```r
# Loading Data pt 2
# btw this sucked
wins_and_losses_2017 <- read.csv("./data/win_loss_2017.csv") %>%
  mutate(split = gsub(" ", "", split, fixed = TRUE), team = gsub(" ", "", team, fixed = TRUE)) %>%
  mutate(split = toupper(split))
```

Preliminary Plots

Comparing The Astros 2017 Home and Away Data

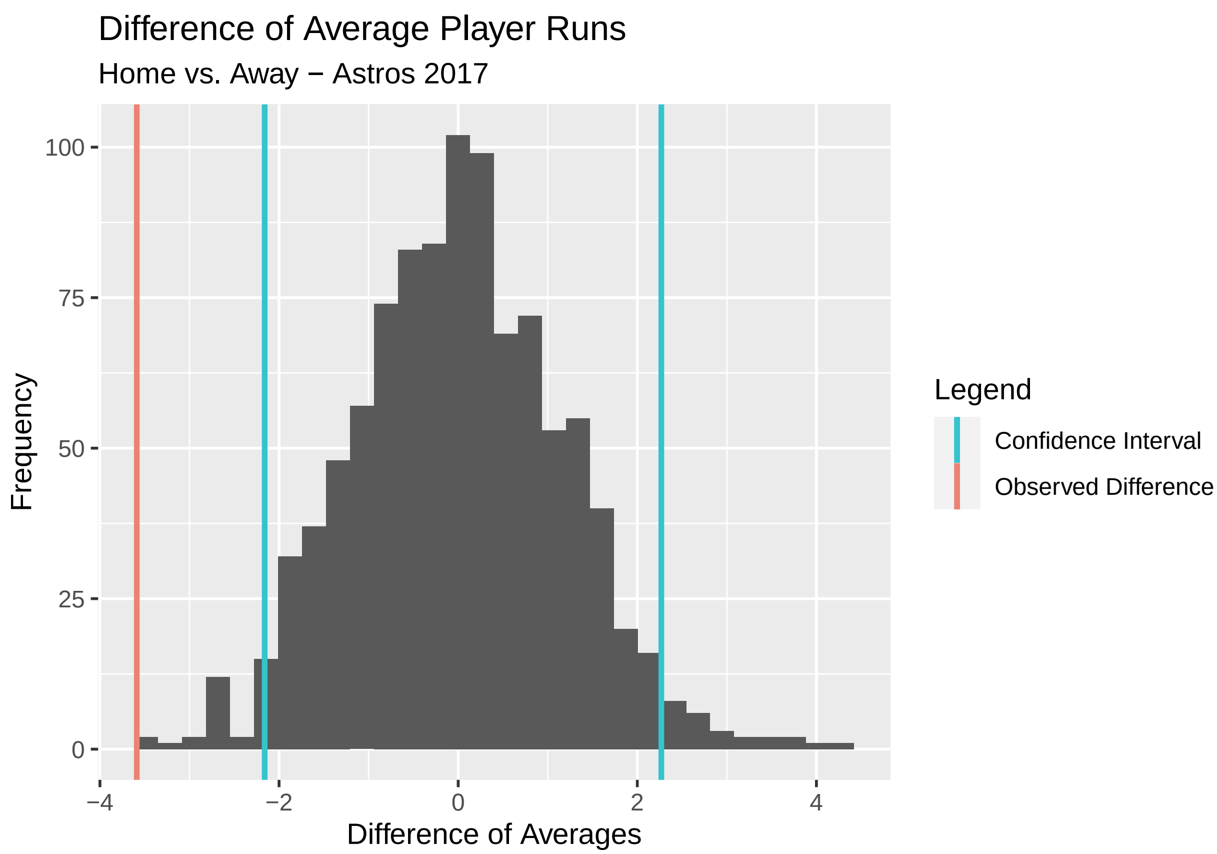
to start we will look at some of the variables we take a look some comparisons of specific variables between home and away games.



These box plots show how there is a clear difference between home and away games. The largest difference is in average Hits (H) per season that a player gets between home and away games.

Runs analysis - Home vs. Away

To do this analysis, we will be doing a monte carlo simulation on the difference of average runs a player gets per season between home and away games. Runs will be modeled by a poisson random variable. Under the assumption that Runs from home and away games comes from the same distribution, we will set lambda as the mean of runs for home and away games combined.



After performing a 95% confidence interval on our data, we see that our observed difference falls outside of the confidence interval, leading us to reject our initial assumption. Thus, we can say with 95% confidence that there is statistical significance in the difference of Run means between home and away games.

However, because Away game Runs are lower than Home game Runs, this implies that the Atros cheating had a positive effect on their performance.

Attempting Regression

Next we will attempt to perform a regression to predict the percentage of games that a team will win based on their summed statistics in Runs, Hits, Home Runs, Runs Batted In, and split. To do this we will have to combine a few datasets: the data we have for specific players from baseball-reference.com and data for team statitcs (games won, lost and played by home and away).

```
summary(lm.2017)
```

```
##
## Call:
## lm(formula = win_percent ~ R + H + HR + RBI + split, data = lm_data_2017)
##
```

```
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.139163 -0.049089 -0.002704  0.041373  0.151519
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.2840020  0.1583572   1.793   0.0792 .
## R            0.0044615  0.0020957   2.129   0.0384 *
## H           -0.0006206  0.0003044  -2.039   0.0470 *
## HR          -0.0003156  0.0007660  -0.412   0.6822
## RBI         -0.0028403  0.0022090  -1.286   0.2047
## splitHOME    0.0487977  0.0186055   2.623   0.0117 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06507 on 48 degrees of freedom
## Multiple R-squared:  0.5357, Adjusted R-squared:  0.4874
## F-statistic: 11.08 on 5 and 48 DF,  p-value: 3.964e-07
```

From our results we can see that our regressions are very bad. None of the p values are below .05 for both home and away games. In the future we will really need to improve upon this model.

Discussion on Progress/Challenges

Challenges:

- On our last project report, the results that we gathered from the data were not what we expected. We originally cleaned the dataset wrong and it affected the outcome, giving us reason to believe the Astros didn't cheat. After going back and changing how we read in the data, our graphs prove what we were originally expecting, and show initial evidence that they were cheating. Our original findings were very surprising because this cheating scandal wouldn't have been a scandal if it weren't for some correlation and evidence that suggest they did. This was a big challenge because after our last project report, we thought we were gonna have to go back and change our whole project. But now we are back on track and can continue the analysis of our questions.
- Another challenge we faced during this project report was trying to figure out how to use a linear model in our analysis. We originally wanted to use a regression model on all of the variables in our dataset to predict wins and then we could see if home games (the supposed cheating) was a statistically significant variable. After initially trying this, we realized this approach was not a very good/strong indication of how home games helped them cheat. Then, we thought it would be beneficial to compare individual Astro players pre-season stats, to that of the regular season, and compare home and away. This also came with its issues though because there wasn't individual player data for preseason. So we moved onto our next thought process
- From the data site Kieth gave us from the last project report, we found a perfect data set split up into the home and away games, but there were no column names for 161 variables. So we used python to assign column names from the reference sheet online.
- In doing this, we found out how hard it is to try and piece together all the datasets. This method also increases the margin of error, which is something we can compute and analyze for the final report.

Progress:

- We came to the realization that some of our data cleaning from our last project report was skewing the results. We fixed that problem, re-fit the previous graphs, and re-analyzed them
- We spend a tremendous amount of time reading in new data and cleaning it, in order to form a regression model. After forming a linear model to predict win_percentage based on Runs, Hits, Runs Batted In, Home Runs, and Split (Home vs, Away), we found that the model sucks, like a lot. So we will need to adapt our model to work better in the future, but we did not have enough time to finish it now.

Next Steps

- We want to figure out why the new regression models we created were so bad. Before the final project is due, we will work on how to improve our model, and attend office hours to gain some insight as why they currently are the way they are. We can then plot the models to visually see how they are doing
- We also might want to look into the performance of teams throughout the league at the Astros' home stadium to see if the conditions under which teams play at that location (called park factors) have any affect on performance, indicating a possible reason for the Astros lower performance at home.
- We are also considering looking into things such as individual Astros players performance and test to see if there were any noticeable, improbable differences between years, comparing it to the change in performance for other teams' top performers.