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

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.

Data for different teams and seasons are also present in the database if needed for further analysis.

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.

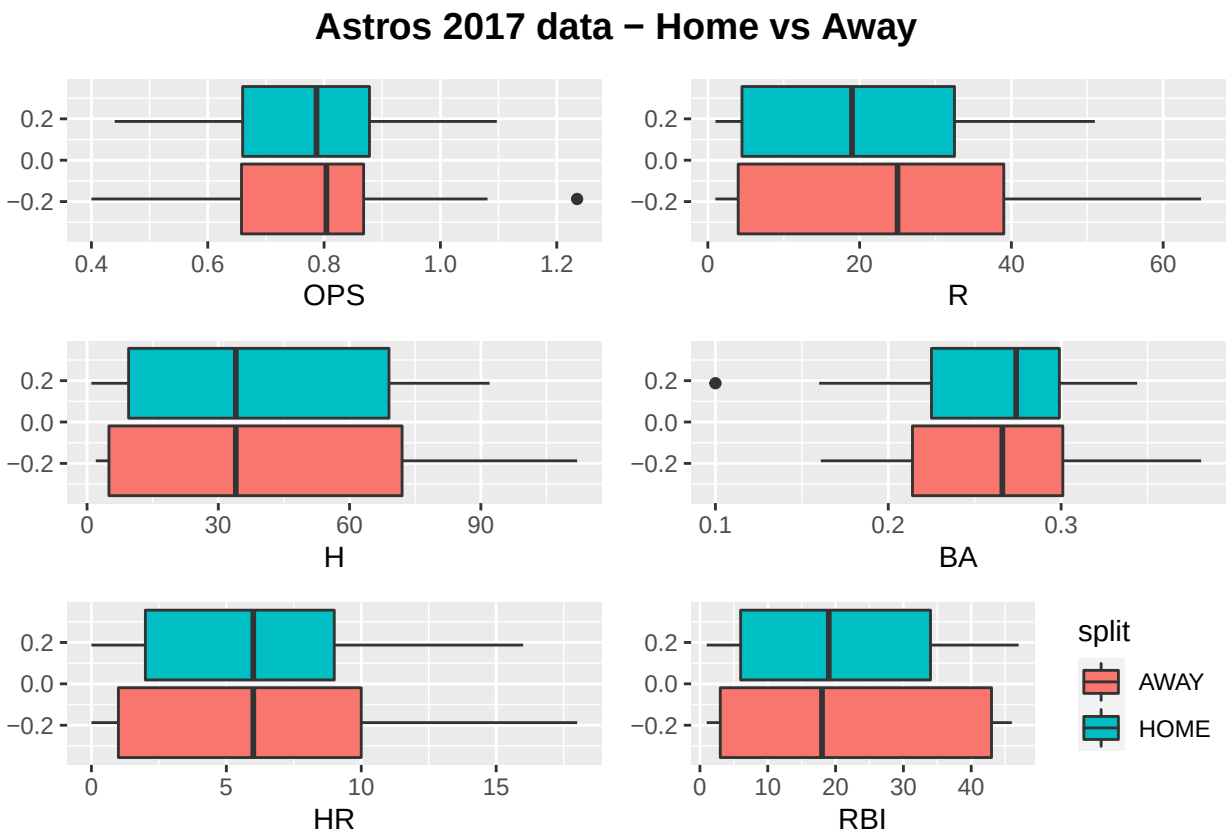
Loading Data

```
astros_home <- read.csv("./data/astros_2017_player_splits_home.csv") %>%
  mutate(Name = str_split_fixed(Name, "\\\\", 2)[,1]) %>%
  mutate(split = "HOME")

astros_away <- read.csv("./data/astros_2017_player_splits_away.csv")%>%
  mutate(Name = str_split_fixed(Name, "\\\\", 2)[,1]) %>%
  mutate(split = "AWAY")

astros_2017 <- rbind(astros_home[-length(astros_home[,1]),], astros_away[-length(astros_away[,1]),]) %>%
  drop_na()
```

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.

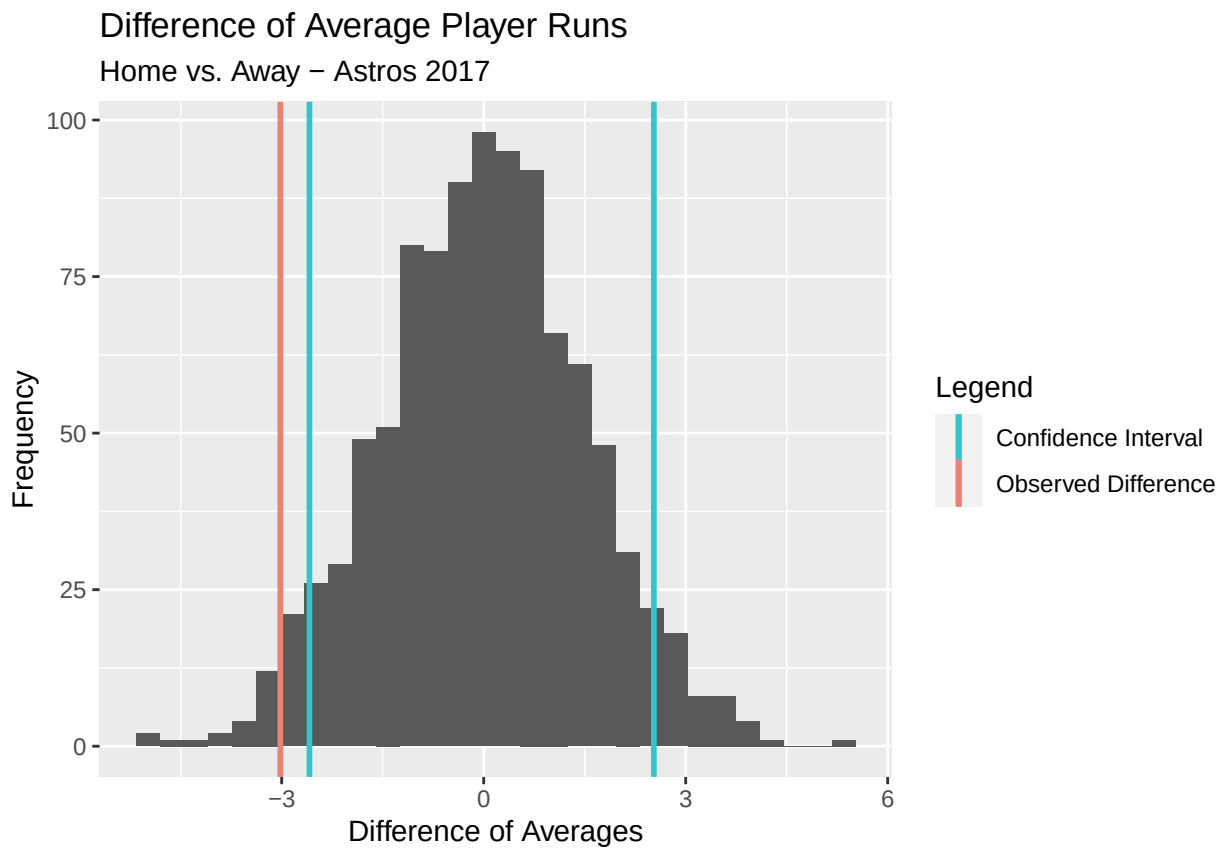


These box plots show that there is not much difference between the home and away games. However there are some patterns that we can explore: large differences between variables, and the variance of the distributions.

The largest difference in means that we can see is the average Runs (R) per season that a player gets between home and away.

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 90% 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 90% confidence that there is statistical significance in the difference of Run means between home and away games.

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

Discussion on Progress/Challenges

Challenges:

- We have found that the data we found for the 2017 Astros was not as detailed as we had hoped. The available dataset from the database that specifies home vs. away games is in form of each players total statistics, not game by game data. If we had game by game home and away data, it would be easier for us to do differnt statistical tests on specific variables and possibly gain more insight during our preliminary analysis.
- Plot 1: *Astros 2017 data - Home vs. Away* - we did a very simple statistical analysis for this plot, and the results seem to suggest that there is no real difference between the home and away statistics, pointing us to believe that the cheating during the 2017 season had no real effect.

- Plot 2: *Difference of Average Player Runs* - when we performed a 95% confidence interval on this monte carlo simulation, the observed difference landed just inside of the confidence interval, which gave us inconclusive results. However, when we ran a 90% confidence interval instead, the observed difference fell outside of the confidence interval, telling us that there was statistical significance. This result supports the opposite of our inital thoery. Thus our problem lies with which percent confidence interval to choose.

Progress:

- So far we have been able to identify that of our variables, Runs, seems to be the only one with statistical significance, and that the cheating could actually have had an adverse affect on the Astros, making them perform worse than if they hadn't cheated.

Next Steps

One of our goals for the remainder of the semester is to use prediction models to answer the statistical question on how the supposed cheating affected individuals play separate from the team. We also planned to answer the statistical question, "how might the season have looked if they didn't cheat", but after finding that cheating had a negative effect on the performance, we may have to alter this statistical question a bit. We also still want to look into the Astros vs. every other team in the league during 2017

-regression?