Holly Ghaemi

M.S. Statistics at the University of Houston

Astro’s R&D Candidate

1. **Describe something interesting or challenging that you learned recently.  How do you learn best?**

This past Thanksgiving Break, I spent more time coding in R than feasting on turkey and mashed potatoes to say the least. My family from Dallas visited Houston and I had the opportunity to collaborate with my cousin, a data scientist, on a project for my linear models course.

While working together, he introduced me to the dplyr package in R. Dplyr is a package in R which offers fast and consistent commands for data manipulation. Essentially, what had previously required 8 lines of redundant code, was reduced to 3 clean and efficient lines after a quick hour of learning how to properly utilize the functions in the package. This experience demonstrated how eminently valuable a knowledgeable mentor who explains concepts, encourage questions and has material-related discussions can be.

Another reason I enjoy working in R and coding is because I am able to apply empirical methods of trial and error. I write a code, run it and am immediately made aware of the level of fruition. Trial and error is an interactive method of learning that effectively engrains techniques into my reasoning and helps me achieve levels of understanding that eventually become second nature.

At first glance, the dplyr package is not intuitive. Therefore, my initial approach was to define functions such as filter(), slice(), etc. After this step of simplification, I was on my way to gaining a deeper understanding of the package. One of my main approaches to a tough problem or concept is breaking it up into easier, more understandable pieces. First, I gain a simple overview. Then, I delve into the details by defining vocabulary words, listing steps of the process and building an organized plan of study. Soon enough, simplified pieces that I’ve established merge cleanly and provide me with a clear and thorough understanding of the problem at hand.

1. **Imagine you are given free range at our organization to investigate any questions related to making our team better. What would you most want to research if you were able to collect the data and/or conduct interviews? How would you go about conducting this research?**

Wouldn’t it be ideal if there was an algorithm that calculated and forecasted the desired traits of the ideal baseball player? One of the challenges in building a better team is considering a pool of talented baseball players and determining which ones are the most outstanding players within that group. To take on this task, it is important to first identify what traits, characteristics and life experiences are correlated with outstanding performance. After establishing these relationships, recruiters can better focus their efforts based on variables that are proven to be significant predictors of player performance.

Collecting data is a clear first step in the process of research investigations. I would create standardized questionnaires for players on the team, interview them and record everything from their heights, weights, and other physical measurements to more personal life experiences such as the marital status of their parents, country of origin, the size of the graduating class of their high schools, the age they began playing baseball, their level of self-perceived attractiveness and whether or not they liked their coaches and teammates. Quantitative body measurements can be indicators of athleticism whereas life experiences can reveal the inner workings of a player’s motivations, self-confidence and resiliency responses. Additionally, from historical archives detailing information about past players, I would try to collect more information to gain a historically broader timeline of perspective.

To ensure that the data is properly formatted and compatible with statistical study, categorical variables would be factorized in R and continuous variables would remain in their numeric formats. As for the response variable of outstanding performance, this will be a numerical and continuous response variable that would be derived from a tailored scoring system that makes use of the available performance statistics of the individual players. After collecting and properly formatting the database, I would run a correlation plot to gain a preliminary idea of relationships between the various explanatory variables and the predicted variable of outstanding performance score. I would further use lasso methods and/or the step() function in R to identify significant predictor variables. Finally, I would run a regression with the refined set of explanatory variables. The final task would be to use the regression to help estimate the performance scores that can be expected from players that are currently on the radar of the Houston Astros as potential recruits.

1. **Given the data in the attached files, determine which of Jose Altuve or Mike Trout was the faster baserunner in 2016. The attached files include player tracking information for balls hit into play during 2016 when the player was either the batter or the runner at first or second base.  The definition of faster speed in this baseball context is up to you to determine and justify.  Construct a chart or other visualization to communicate your findings.  Please include in the form of a Jupyter, RMarkdown, or similar notebook your R or Python code used to analyze the data, produce the visualization.**

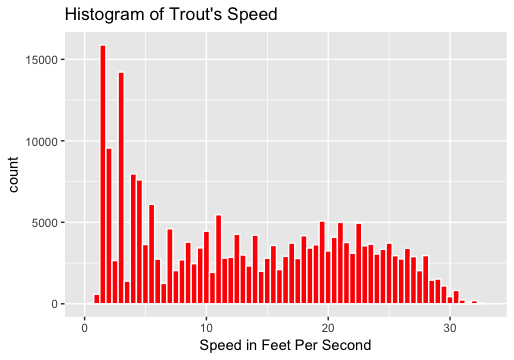
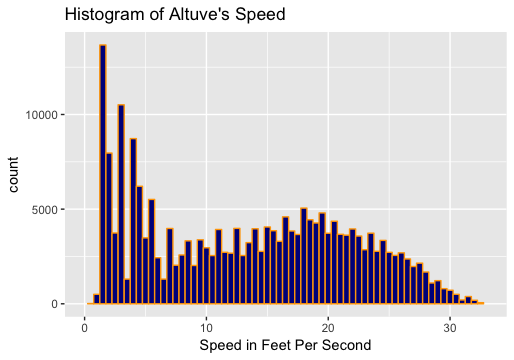
The x, y coordinates of both Altuve and Trout with their respective times gives rise to the following definition of velocity, aka speed, being measured in feet per second:

The following is an overview of the results found in R:

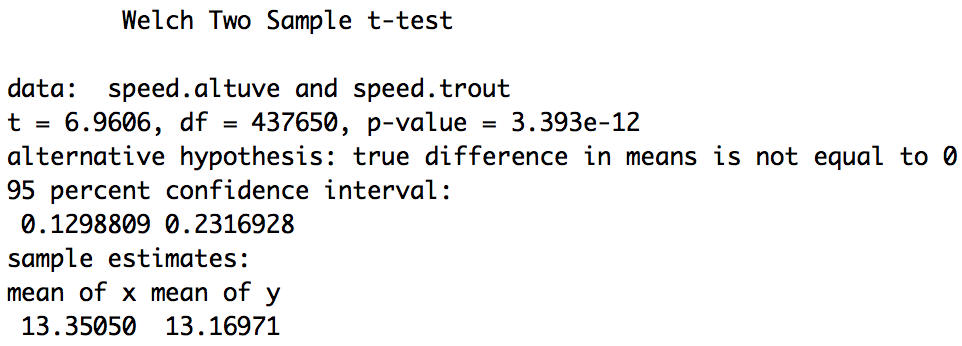
José Altuve Mike Trout

|  |  |  |
| --- | --- | --- |
| José Altuve (feet/sec) | Statistic Type | Mike Trout (feet/sec) |
| 43.38 | Max Speed | 45.33 |
| 13.35 | Average Speed | 13.17 |
| 13.68 | Median | 12.54 |
| 8.44 | Standard Deviation of Speeds | 8.75 |
| 13.31 | Lower Confidence Interval | 13.13 |
| 13.39 | Upper Confidence Interval | 13.21 |
| 20.62 | Mode Speed > 7 ft/sec | 29.23 |
| 379 | Frequency of Mode Value | 399 |



A two-sample t-test on the resulting calculated vectors of Altuve and Trout’s speeds was run to ensure the means can be considered statistically different. The null hypothesis was that the mean speeds of both players are equal, and the alternative hypothesis was that the mean speeds of the players are not equal. The following results were obtained:



The p-value is much smaller than the alpha level of 0.05 at p-value = 3.39e-12. Therefore, we can reject the null hypothesis and state that the means of Altuve and Trout’s speeds are indeed different. Therefore, we can continue with our comparison of their point estimates.

Based on these statistics and histograms, it is clear that Mike Trout’s maximum and mode speeds are greater than that of José Altuve. However, the average speed, median speed and confidence intervals of Altuve’s speed are higher. Additionally, the standard deviation of Altuve’s speed is lower. These statistics show that although, Mike Trout was able to achieve a higher max and mode speed in 2016, José Altuve demonstrated a narrower spread of speeds, indicating that his speed is more consistent. You can also see this conclusion when observing Trout’s histogram which shows a less normal distribution than that of Altuve. The spread of Trout’s speeds more closely resembles the uniform distribution indicating that his probability of attaining varying speeds, such as an approximate 12 feet/second and 22 feet/second, were roughly equal in 2016. Therefore, I would state that we can expect José Altuve to be on average, faster than Trout.

However, other confounding variables that were not controlled and could possibly affect speed should also be considered when reviewing these results. For example, the number of home runs a player hits can affect how likely they are to sprint to bases as it is unnecessary to run quickly if a clear home run is hit. This and other factors should be more closely examined in the final conclusions of speed analyses.